Journal of the Audio Engineering Society
Vol. 63, No. 5, May 2015 (© 2015)
DOI: http://dx.doi.org/10.17743/jaes.2015.0021

Autonomous Multitrack Equalization Based on Masking Reduction

SINA HAFEZI AND JOSHUA D. REISS, *AES Member*
(sina.clamet@gmail.com) (joshua.reiss@qmul.ac.uk)

Queen Mary University of London, London, UK

Spectral masking is when the threshold of audibility for one sound is raised by the simultaneous presence of another sound. In multitrack music production, this results in less ability to fully hear and distinguish the sound sources in the mix. We design a simplified measure of masking based on best practices in sound engineering. We implement both off-line and real-time, low latency autonomous multitrack equalization systems to reduce masking in multitrack audio. We perform objective measurement of the spectral masking in the resultant mixes and conduct a listening test for subjective comparison between the mix results of different implementations of our system, a raw mix, and manual mixes made by an amateur and a professional mix engineer. The results show that autonomous systems reduce both the perceived masking and objective spectral masking and improve the overall quality of the mix. We show that our offline semi-autonomous system is capable of improving the raw mix better than an amateur and close to a professional mix by simply controlling one user parameter. Our results also suggest that existing objective measures of masking are ill-suited for quantifying perceived masking in multitrack musical audio.

1 INTRODUCTION

In sound engineering and recording, mixing is the process of combining multiple recorded sounds, referred to as "multitrack," into one track known as a "mix down." In the process of mixing, the source signals' level, frequency content, dynamics, and panoramic position are manipulated, and effects such as reverberation may be added for artistic reasons in order to make a mix more enjoyable as well as for technical reasons to correct problems coming from poor recording, performance, orchestration, etc.

Masking is defined as the process by which the threshold of audibility for one sound (the maskee) is raised by the presence of another sound (the masker) [1–3]. There are two main types of auditory masking: *Spectral Masking*, also known as simultaneous masking or frequency masking, occurs in the frequency domain, and *Temporal Masking*, also known as non-simultaneous masking, occurs in the time domain. In this research we only focus on spectral masking in multitrack mixing and will refer to this phenomenon as simply "masking."

The amount of masking will vary depending on characteristics of both the maskee and the masker and will also be specific to an individual listener. When multitrack audio is mixed, masking reduces the listener's ability to distinguish the sound sources [4–6]. This makes the mix confusing, underwhelming, and unclear.

Gonzalez and Reiss [7] addressed the issue of masking by providing a system that adjusts the levels of tracks with overlapping frequency content in order to reduce masking of a target track. This approach, though using a measure of masking similar to the one provided herein, applies gain changes, not equalization. Thus it applies quite harsh changes across the entire frequency range and, as noted in [8], may lead to a reduction in the overall dynamic range of the mix. Furthermore, [7] was aimed only at reducing masking of the target track and not reducing masking of the overall mix.

Audio engineers employ three main tools for reducing masking in multitrack mixing [5, 6, 9], the first two having been implemented in intelligent systems: adjusting the relative level of each track (as in [8, 10–12]), panning the tracks that cause masking to different spatial positions (as in [13–15]), and equalization of tracks.

Equalization, or EQ, involves the use of linear filters with adjustable parameters to manipulate the frequency content of audio signals. Equalization of tracks may be used creatively, but in the context of masking reduction it can be applied to ensure that each track dominates only a portion of the frequency domain and to avoid the strong overlap of frequency content from multiple sources.

In [16] an approach to automatic multitrack equalization was proposed 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 [17], and their approach does not directly address spectral masking.

In this paper we derive an autonomous system that applies equalization to all input tracks in order to reduce masking in the resultant mixdown. Such a system is a content-based equalizer and falls under the category of Cross-Adaptive Audio Effects (XA-DAFx) [18]. The main idea behind an Adaptive Digital Audio Effect (A-DAFx) is that the processing applied by the effect on the input is controlled by the analysis of sound features derived from the input. Additionally, an XA-DAFx is defined to be a type of A-DAFx that is multi-input multi-output (MIMO) and the individual processing of each input depends on the content of all inputs.

The paper is structured as follows. Sec. 2 introduces the simplified masking measure used to develop autonomous multitrack EQ systems, based on algorithmic implementations of manual approaches described in the literature. Sec. 3 describes the structure of the systems that we constructed. This is split into discussion of their analysis and their processing stages. In Sec. 4 we describe the implementations of such systems, including both off-line and real-time approaches. This section also provides the full details and justifications for the parameter settings chosen in each implementation. Sec. 5 describes objective evaluation of our systems using measures of masking from psychoacoustics research, as opposed to those from sound engineering practice. In Sec. 6, the subjective evaluation of our implementations is described and the results of this evaluation are depicted. Finally, Sec. 7 concludes with a discussion of the implications of this work.

2 MASKING MODEL

The widely accepted model of masking proposed by Moore [1] is based on extensive psychoacoustic experiments, especially those described in [3]. In this model, the excitation patterns for the two sounds are calculated first. The excitation is meant to correspond to the average neural activity in response to a steady sound as a function of frequency and is calculated as the squared sum of the output of each auditory filter as a function of the filter's center frequency. Then, the regions with significant excitation overlap in time and frequency are detected, and finally a decision is made for each of these regions in which a sound is labelled as masker and the other one as maskee.

Based on this auditory model, masking in multitrack audio has been quantified with a masked-to-unmasked ratio [19], a cross-adaptive signal to masker ratio [20], and measures of partial loudness in a mix [10, 21]. However, these metrics are computationally intensive and do not easily lend themselves to use in a masking reduction system, especially if deployed for real-time use. Furthermore, only [21] provided formal evaluation against human perception with real world signals, musical or otherwise. In fact, the evaluation performed in [21], as well as informal evaluation described in [10], suggested that the auditory model of masking yielded highly inaccurate results when applied to multitrack musical audio. Therefore we aim to design and assess an alternative measure of masking that is inspired by best practices in audio engineering and suitable for deployment in a real-time multitrack equalization system.

Most of the approaches [5, 6, 9] to manual multitrack equalization proposed by professional sound engineers are based on a specific instrumentation, are constrained to unique properties of the discussed case, and leave some analysis and processing tasks to personal artistic taste and interest. Although individual factors and taste make the automation of equalization difficult, some similarities and shared points in these approaches allow us to develop a general definition and algorithm for masking reduction in musical context. The key and common points are as follows:

- It is more reliable to attenuate the masked frequency regions instead of boosting the unmasked frequency regions [6, 9]. Although [17] found that expert mixers do not tend to cut more than boost, the masked frequency regions are most likely smaller in comparison to the unmasked frequency regions. Therefore attenuating the masked regions has less impact on the balance between the loudness of tracks. Also a boost on one track can be achieved by attenuation of the masking tracks (mirror equalization) [5, 6].
- The frequency spectrum can be divided into *essential* and *nonessential* frequency regions. The essential regions are most likely the highest amplitude portions of spectrum and nonessential regions are most likely the frequency regions that are easy to attenuate with low impact on timbre change and loudness balance between the tracks [5, 6, 9].
- For a given track, the frequency regions that are mainly covered by other tracks can be attenuated [5, 6, 9, 17].

Our measure of masking in multitrack audio is directly based on these manual multitrack equalization approaches and, hence, does not explicitly incorporate auditory models. However, in Sec. 5 it is compared with measures based partly on auditory models.

In our model, masking occurs at a given frequency region if both of the following conditions are met:

(1) The magnitude of the masker is higher than the magnitude of the maskee in that frequency region;
(2) That frequency region is nonessential for the masker and essential for the maskee.

For condition (1), it should be noted that a masker with less magnitude than maskee still raises the threshold of audibility for the maskee in that frequency region and may cause masking, although this case is not considered in our analysis and treatment.

Condition (2) is important because we want to apply masking treatment with low impact on perceived timbre. Absence of this condition would result in a situation where in every frequency region there is always a track with largest

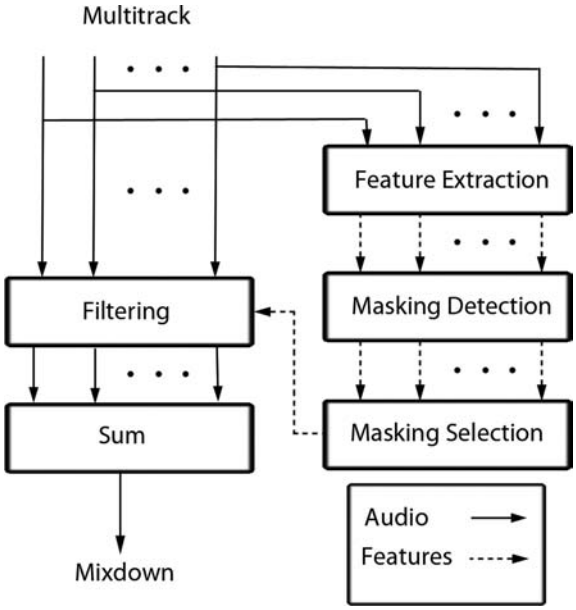

Fig. 1. Block diagram of the system.

magnitude, and hence we would always have masking in all frequencies.

The frequency regions (bins) are ranked based on their magnitude so that rank 1 has the highest magnitude among all bins. In our model, 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 Eq. (1),

$$M_{AB}(f,t) = \begin{cases} X_A(f,t) - X_B(f,t) & if\ R_B(f,t) \le R_T < R_A(f,t) \\ 0 & else \end{cases} \quad (1)$$

where $X_I(f,t)$ and $R_I(f,t)$ are respectively the magnitude in decibels and the rank of frequency f, at time t for track I. R_T is the maximum rank for a frequency region to be considered essential. This equation provides a formal mathematical description of conditions (1) and (2) above. If $M_{AB}(f,t)$, referred to simply as M for brevity, is greater than zero, then f is considered a dominant nonessential frequency, i.e., at frequency f, the masker dominates over the maskee (condition (1)) but this frequency is considered nonessential to the masker (condition (2)).

We reduce masking by attenuating the masker by the value M, over a range of frequencies centered at the dominant nonessential frequency. R_T and M are set differently based on the implementation method (see Sec. 4).

3 SYSTEM

Our system, shown in Fig. 1, determines the essential and nonessential frequency regions of each track, reports the positive values of M in Eq. (1), finalizes the amount of attenuation for masker frequencies of each track, and sends the frequency and amount of attenuation to the equalizers dedicated to each track. In order to focus on EQ and exclude other types of masking treatment, we assume that the relative loudness level of tracks of the multitrack are properly adjusted, the masking determination is performed on a monaural-converted copy of the tracks, and the same equalization will be applied to left and right channels of any stereo tracks.

The system is divided into two main parts known as Analysis and Processing. Full details are specific to each of the implementations and described in Sec. 4. However, each implementation shares the same general framework described in this section.

3.1 Analysis

This part handles the detection and measurement of the spectral location and amount of masking for each track. It consists of three types of operation block.

The *Feature Extraction* block calculates the magnitude of each frequency region of the input track and ranks the regions based on their magnitude.

In the *Masking Detection* block, each input track is considered as potential masker and all other tracks as potential maskees. For K tracks there will *be* $K(K\text{-}1)$ pairs that will be considered and analyzed. We may find multiple positive values of M for a particular frequency region among different pairs, meaning that the input track may mask multiple tracks at the same frequency. In this case we only consider the maximum value of M among all the detected values for that frequency region.

The *Masking Selection* block determines which frequencies will be equalized. If the number of detected masking occurrences is greater than the number of filters in our equaliser, we give priority to the highest values of M.

3.2 Processing

This part consists of an EQ per track followed by a mixer that sums the output of each equalizer into a mixdown channel. The results of the analysis part are used as the filter parameters of the equalizers. The center frequencies and attenuations of the peaking filters in the EQ are respectively the frequencies and the values of M in Eq. (1), which were selected in the Masking Selection block.

4 IMPLEMENTATION

Four different implementations of our system were made. As shown in Table 1, these are distinguished by their run type, degree of automation, and parameter constraints.

4.1 Offline Implementation

The offline versions are time-invariant since the equalization settings remain constant over time and non-causal since the equalization at any time depends on the past, present, and future. These implementations analyze the average masking occurrence on each track and apply a constant EQ per track over the entire duration.

Two versions for offline implementation were made. One is fully autonomous and the other is semi-autonomous in order to give the user control over the strength of

Table 1. Specifications of implementations.

Name	Graph Label	Run Type	Autonomy	Constraints
Offline Fully	OfF	Offline	Fully	–
Offline Semi	OfS	Offline	Semi	–
Real-time Unconstrained	OnU	Real-time	Fully	–
Real-time Constrained	OnC	Real-time	Fully	Filters' Gain and Q

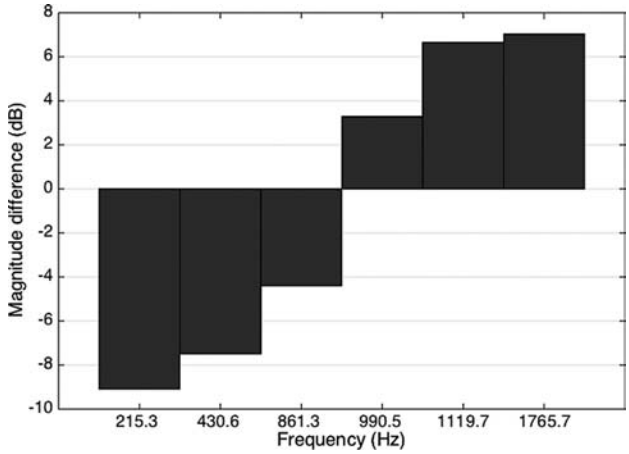

Fig. 2. This shows the difference in magnitude (M value) between the horn (maskee) and cello (masker) tracks at those frequencies that are essential for the horn but nonessential for the masker. Positive values imply that masking reduction is needed.

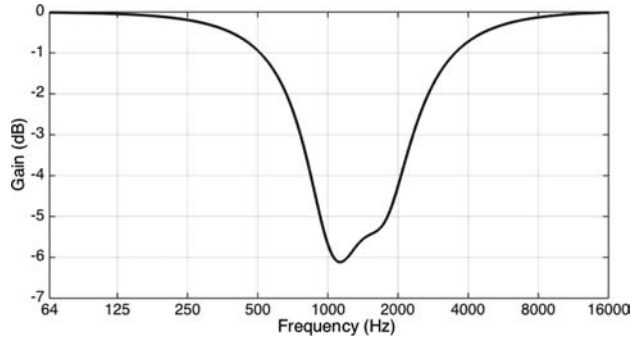

Fig. 3. The fully autonomous ($S = 0$) equalization filter applied to the cello track in order to reduce masking of the horn track.

equalization. Both versions have the same analysis part but differ in processing (equalization).

We use an FFT to obtain the magnitude of each frequency region in a track. The spectrum of each track is obtained by averaging the results of FFTs on non-overlapping 1024-point frames over the entire length of each track. We then use Eq. (1), with $R_T = 10$ based on an informal listening test, to classify the peaks of the magnitude response into essential and non-essential frequency bins.

The equalizers consist of three second-order IIR peaking filters in series. So after passing the averaged spectra into the Masking Detection blocks, a maximum of three masking occurrences with the highest value of M per track are selected for equalization.

The frequency of the selected M value is assigned to the filter's center frequency. The filter's Quality Factor, Q, is set to 2, also based on informal listening tests. The fully and semi-autonomous versions use B in Eq. (2) as the filter's gain.

$$B = -2^S M, \tag{2}$$

where S (or Strength) is a user parameter in the semi-autonomous version to scale the amount of attenuation. For the fully autonomous version $S = 0$, so that the system attenuates as much as it detects. Whereas in the semi-autonomous version, for positive or negative values of S, the user adjusts S to positive or negative values in order to respectively scale up or down the amount of attenuation.

In the semi-autonomous approach, the system first finishes the analysis part and applies equalization with $S = 0$. The user may then adjust S in order to find the best sounding output.

In the mixdown stage, all the processed tracks are summed and the mixdown track is normalized such that the peak amplitude is 1 in order to avoid clipping.

Figs. 2 to 4 illustrate this implementation for the case of two eight-second tracks, horn and cello, where the horn track is masked by the cello track. Six frequency bins are identified as essential for the horn but nonessential for the cello, using Eq. (1) with $R_T = 10$. The magnitude difference M between the masker cello and maskee horn tracks at each of these frequencies is depicted in Fig. 2. The positive values of M indicate frequencies where masking reduction should be applied. Fig. 3 shows the equalization filter applied to the cello, which consists of three notch filters in series, with $Q = 2$, center frequencies set to those considered essential for the maskee and where the masker dominates over the maskee, and gain set to the respective M values. Finally, Fig. 4 shows the spectrum of the masker before equalization and after equalization for both fully autonomous ($S = 0$) and semi-autonomous ($S = 2$) cases.

4.2 Real-Time Implementation

The real-time versions are implemented in C++ as a 10 track, stereo-input VST audio effect plug-in and can be used in any host application that supports MIMO (multi-input multi-output) VST. The plug-ins are time-variant since the equalization settings vary over time and causal since the equalizations at any time depend on the past and present only.

The real-time versions operate on a frame-by-frame basis. For each incoming audio frame, the system calculates

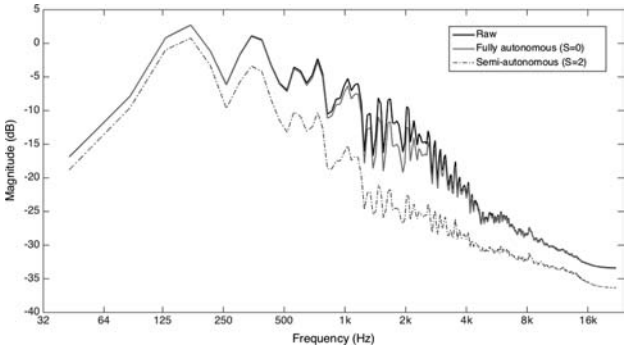

Fig. 4. The spectrum of the cello track (masker) before and after equalization for fully autonomous (S = 0) and semi-autonomous (S = 2).

M as defined in Eq. (1), detects and selects the masking occurrences, smooths the decisions using an exponential moving average (EMA) filter, and applies a time-varying EQ on each track in the processing stage. As opposed to the more computationally expensive FFT used for offline implementation, a filter bank approach was used here to calculate the magnitude response. Thus frequency resolution was sacrificed to ensure real-time operation when analyzing multitrack audio. Use of the filter bank and EMA filters also allowed us to minimize latency, thus ensuring that the plug-in could be used in a live sound mixing environment.

The filter bank consists of multiple single-channel second order Butterworth bandpass filters set up in parallel, each centered at a fixed frequency value. A monaural-converted copy of the signal goes through each bandpass filter. For a given filter, the center frequency represents the frequency band and, as shown in Eq. (3), the RMS of the filtered signal represents the magnitude of that band,

$$X(f) = RMS((x * h)[n]) = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left(\sum_{m=-\infty}^{\infty} x[m]\cdot h[m-n]\right)^2}, \qquad (3)$$

where *X(f)* is the Root-Mean-Square (RMS) of the input signal after being filtered by the bandpass digital filter centered at frequency *f, h* is the impulse response of that bandpass filter, *n* is the sample index, *x* is the input digital signal, (*x***h*)[*n*] is the convolution of the signal and filter impulse response, and *N* is the length of the input signal *x*. The following ISO standard octave-band center frequencies [22] were used: F_c = [31.5, 63, 125, 250, 500, 1K, 2K, 4K, 8K, 16K] Hz.

Having obtained the magnitude response of the tracks, frequency ranking is performed. As we only have ten frequency bands, an informal listening test was performed and the three frequency bands with the highest magnitudes were selected as essential ($R_T = 3$).

Masking Detection and Selection is performed in the same manner as discussed in Sec. 3.

The processing stage consists of five second order Butterworth peaking filters in series per track. Therefore only a maximum of five masking occurrences with the highest values of *M* are selected in the Masking Selection block.

Fig. 5. User Interface of the real-time unconstrained VST plug-in. In the Filters section, the user can see the EQ settings of the selected track. The plug-in outputs the mixdown but the user is also able to solo an individual track. The speed slider controls the time constant τ. The Pre-amp slider changes the volume of the overall multitrack before analysis. Bypass and Reset buttons respectively deactivate and reset the plug-in.

Since our EQs are time-variant and the filters' parameters (center frequency and gain) need to be updated smoothly, we apply an EMA as a smoothing function on the filter's center frequency and attenuation *M*. EMA is a first-order IIR filter with the difference equation shown in Eq. (4).

$$y[n] = \begin{cases} x[n] & n = 0 \\ (1-\alpha)x[n] + \alpha y[n-1] & n > 0 \end{cases} \qquad (4)$$

where *n* is the sample index, *y* is the smoothed parameter, *x* is the unsmoothed parameter, and α is the smoothness factor between 0 to 1. The closer to 1 α is, the smoother *y* will vary. α is defined in Eq. (5).

$$\alpha = e^{-1/(\tau f_s)} \qquad (5)$$

where f_s is the sampling rate, and τ is a time constant with default value of two seconds.

The center frequency of the filter is the frequency of the selected *M* value and the gain is calculated using Eq. (3) with $S = 0$. Two real-time implementations were designed. The unconstrained version uses $Q = 1$ and has no limitation on the amount of attenuation, whereas the constrained version is more conservative in operation. It sets *Q* to 5 and is allowed to attenuate by a maximum of 6 dB to avoid harsh filtering. For both these versions *Q* stays constant over time.

Fig. 5 illustrates the user interface of the unconstrained real-time implementation.

Table 2. Summary of multitrack songs used in evaluation.

No.	Song	Artist/Band	Genre	No. Tracks	Vocal	Duration (Seconds)	Group	Separate Drums
1	The Road Ahead	Timo Carlier	Acoustic	7	Yes	19	2	Yes
2	Heart Peripheral	AM Contra	Dance	4	No	15	1	No
3	We Feel Alright	Girls Under Glass	Electronic	8	No	30	2	Yes
4	Knockout	M.E.R.C. Music	Hip Hop	9	No	24	1	Yes
5	All That Jazz	Catherine Zeta Jones	Jazz	9	Yes	28	1	No
6	Stan	Eminem	Rap	7	Yes	48	1	No
7	Careless Whisper	George Michael	Pop	10	Yes	50	2	No
8	Feeling Good	Muse	Rock	7	Yes	35	2	No

Generally, real-time systems have a main function that is called for each incoming frame. In our system the input audio frame will be a chunk of all tracks of the multitrack, with adjustable duration that is set by the user from the host application. Our main function, "processBlock," has the pseudo-code shown in the Appendix, and follows the structure given in Fig. 1.

5 OBJECTIVE EVALUATION

A quantitive measure of masking based on the Masked-to-Unmasked Ratio (*MUR*) [19] is used for the objective evaluation. This was chosen over the alternative metrics in [10, 20, 21] since it is the only measure of masking of a track in a multitrack mix that provides a single value and does not require manual customization *MUR* may be defined as (note that the notation here is slightly different from [19]),

$$MUR(x, y) = \frac{L_P(x, y)}{L(x)} \cdot 100\%, \quad (6)$$

where $L_P(x,y)$ is the overall total loudness of a signal x in the presence of a masker y (i.e., partial loudness), and $L(x)$ describes the overall total loudness of the same signal x when the masker is assumed not to be present. Both $L_P(x,y)$ and $L(x)$ are singular values based on averaging over all frames. The loudness and partial loudness are based on the time varying loudness model of Moore, Glasberg, and Baer [23–26] and their calculation, as performed herein, is summarized in [21]. The value of *MUR* ranges from 0 to 100%, where 100 indicates no masking and 0 indicates the signal is completely masked by other sounds. To find an average *MUR* for a multitrack composed of K tracks, x_1, x_2, ... x_K we consider the average of the *MUR* for each track as masked by the sum of all other tracks in the mix,

$$MUR_{Avg} = \frac{1}{K} \sum_{i=1}^{K} \frac{L_P\left(x_i, \sum_{j=1, j\neq i}^{K} x_j\right)}{L(x_i)} \cdot 100\%. \quad (7)$$

Eight songs from a multitrack testbed [27] with varying genres, instrumentations, and number of tracks, shown in Table 2, were used for the objective evaluation. For each song we have five mixes: the simple sum of input tracks without any equalization ("Raw") and the four implementations of our system shown in Table 1. Table 3 contains the average *MUR* for each mix of each multitrack. Each row and column respectively represents the song and the mix. The last column gives the mean improvement in MUR_{Avg}, as a percentage, from the value for the Raw mix for each of our four implementations, i.e., for implementation I, [$MUR_{Avg}(I)$ – MUR_{Avg} (*Raw*)] / MUR_{Avg} (*Raw*), again averaged over all songs.

Table 3. Average masked-to-unmasked ratio in percentage (%).

Song ID	Raw	OfF	OfS	OnC	OnU
1	12.373	12.825	12.399	12.515	12.587
2	28.101	28.233	28.466	27.157	26.898
3	16.985	17.309	18.319	17.080	18.585
4	13.599	13.700	14.521	13.921	14.571
5	12.786	12.968	13.278	12.862	13.212
6	17.631	17.712	17.756	18.223	18.327
7	11.549	11.530	11.807	11.688	11.615
8	14.524	13.982	13.982	14.903	16.244
Mean improvement		0.6%	2.4%	1.1%	4.2%

Based on the MUR_{Avg} metric, all implementations have been successful in the reduction of masking. The most masking reduction was achieved by "Real-time Unconstrained," which is not constrained in the amount of attenuation. As expected, the "Offline Semi" reduced the masking more than "Offline Fully" because the semi-autonomous version gave the user the ability to improve the results of the fully autonomous approach. However, the masking improvements are very minor in all cases. This further validates the need for a listening test in order to examine whether a masking reduction may be perceived.

6 SUBJECTIVE EVALUATION

6.1 Test Procedure

A listening test was conducted to evaluate our implementations. The same eight songs that were used for objective evaluation were used in the listening test. Each test question consisted of comparing seven mixes: the simple sum of input tracks without any equalization ("Raw"), manual equalization done by a professional sound engineer ("Professional"), manual equalization done by a musician possessing amateur mixing skills ("Amateur"), and mixes from each of our four implemented systems. The semi-autonomous version of our system was controlled by the person who did the "Amateur" mix.

In order to reduce the total duration of the test and avoid listener fatigue we only selected a short portion of the music for each multitrack ranging from 15 to 50 seconds.

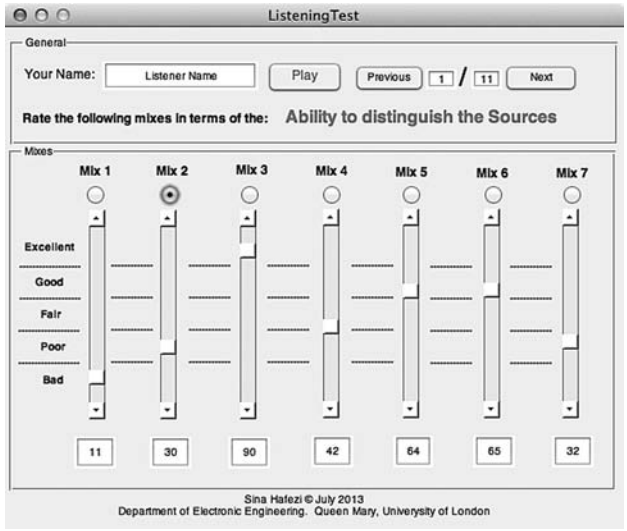

Fig. 6. User Interface of the listening test application.

The songs were divided into two groups, where each group is used for rating mixes in one of the following tasks:

1. Ability to distinguish the sources in the mix;
2. Overall preference for the quality of the mix.

The two different tasks respectively help us in answering the two following questions:

1. How well do our systems reduce the masking in music?
2. How well does the result of our systems satisfy the listener in terms of the general quality?

A listening test application with graphical user interface was designed and implemented for the test, as shown in Fig. 6. The test used multistimulus rating, similar to the MUSHRA framework [28] in which each audio sample is rated from 0 to 100 split up into five descriptors: "Bad," "Poor," "Fair," "Good," and "Excellent." However, unlike MUSHRA, there is no reference, and the raw sum mix may not provide a clear anchor. Thus, the participants were asked to rate at least one mix above 80 and at least one mix below 20, effectively treating one mix as a hidden reference and treating another as a hidden low anchor, as in [11]. This ensures that test subjects use the entire rating scale, but may result in exaggerating the importance of perceived differences. The application notifies the user if this rating condition is not met. For a given case (a multitrack song), the music player loops the song with the selected mix while the user can instantly switch and listen to a different mix of the song by selecting the associated radio button for the mix.

To exclude the effect of perceptual loudness of different mixes on the rating, for a given song, the application normalizes the loudness of all mixes using the ITU/EBU loudness model [29] so that the mixes have the maximum possible equal loudness that avoids clipping.

For each participant, the order of the songs and mixes was randomly changed. The test was run in a quiet, acoustically isolated room (the Listening Room at Queen Mary University of London's Performance Space) under controlled conditions using a professional M-Audio Studiophile Q-40 mixing headphone.

Although we gathered information about the mixing and musical activity background of participants, the listening test application also measures the listening skill of our participants. For Song No. 8, the preferred value of strength S for "Offline Semi," chosen by the user, is zero, meaning that the "Offline Semi" mix is identical to the "Offline Fully" mix. The absolute difference between the ratings of these two mixes for that particular song, named as "distinguishability error," is used to measure the listening skill of a participant. The higher the difference is, the more error the participant had in rating two identical mixes.

A total of 11 subjects between the ages of 20 and 42, all with normal hearing, participated in the test (9 male, 2 female; 9 had listening test experience; 9 experienced in making music; 7 experienced in mixing; 8 familiar with masking).

Table 4 contains the duration of the test and the distinguishability error for each participant. The results from two of the participants with distinguishability error of 30 or above were removed from evaluation.

Table 4. Information on individual participants.

Participant ID	Test Duration (Minute.Second)	Distinguishability Error (/100)
01	27.37	17
02	33.55	15
03	24.50	19
04	19.30	0
05	31.11	35
06	65.05	23
07	46.04	13
08	42.58	2
09	48.33	2
10	56.41	25
11	29.44	34

6.2 Results

6.2.1 Ability to Distinguish the Sources

In this task the participants were asked to rate the mixes in terms of the ability to distinguish the sources. Four songs of different genres and instrumentations were selected. Fig. 7 shows the standard deviation (black vertical line), mean (black horizontal line), and 85% confidence interval (grey box) of the ratings among all participants for each song and each mix.

Fig. 7(a) shows that the "Amateur" mix is mostly rated between "Raw" and "Professional." The highest variety of opinion occurs for "All That Jazz." This song had multiple similar sounding brass instruments, which may cause difficulty in source distinguishability. On the other hand,

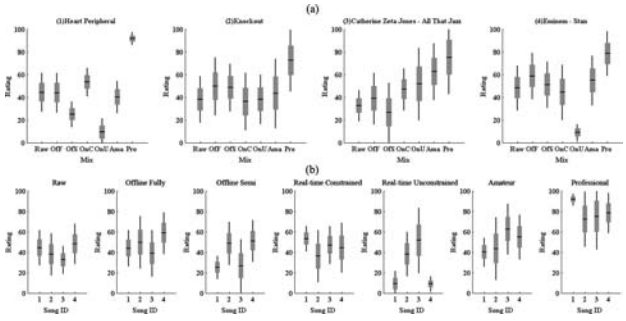

Fig. 7. Task 1: Ability to distinguish the sources. (a) The ratings of mixes for each song. (b) The ratings of songs for each mix.

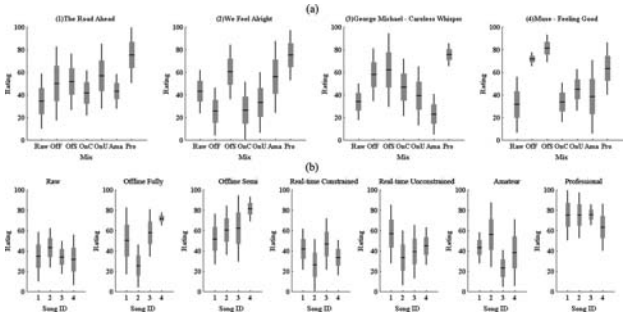

Fig. 8. Task 2: Overall preference. (a) The ratings of mixes for each song. (b) The ratings of songs for each mix.

we have the least variance of rating for "Heart Peripheral," which may be due to the low number of tracks, suggesting that the simplicity of instrumentation may directly affect the simplicity of evaluation and similarity of opinions.

As we can see from Fig. 7(b), "Real-time Unconstrained" has high variance in the ratings. In some cases, such as "All That Jazz," unrestricted attenuation can be an advantage and results in noticeable improvement whereas in "Heart Peripheral" and "Stan," it can result in a worse mix than the "Raw" mix. Comparing "Offline Semi" and "Amateur," which were done by the same person possessing amateur mixing skill, we can see "Offline Semi" is not generally rated higher than "Amateur." This could be due to the fact that the amateur might subconsciously perform mixing in terms of overall preference, whereas here we asked the participants to rate in terms of source distinguishability. Another reason could be the challenging nature of estimating source distinguishability in the mix.

6.2.2 Overall Preference

In this task the participants were asked to rate the mixes in terms of their overall preference. Fig. 8 shows the standard deviation (black vertical line), mean (black horizontal line), and 85% confidence interval (grey box) of the ratings among all participants for each song and each mix.

From Fig. 8(a) we have at least one version of our systems that has the top rating after "Professional." As with Task 1, in Task 2 the participants mostly rated "Real-time Constrained" close to "Raw" mix, which is due to the careful equalization caused by limiting the maximum amount

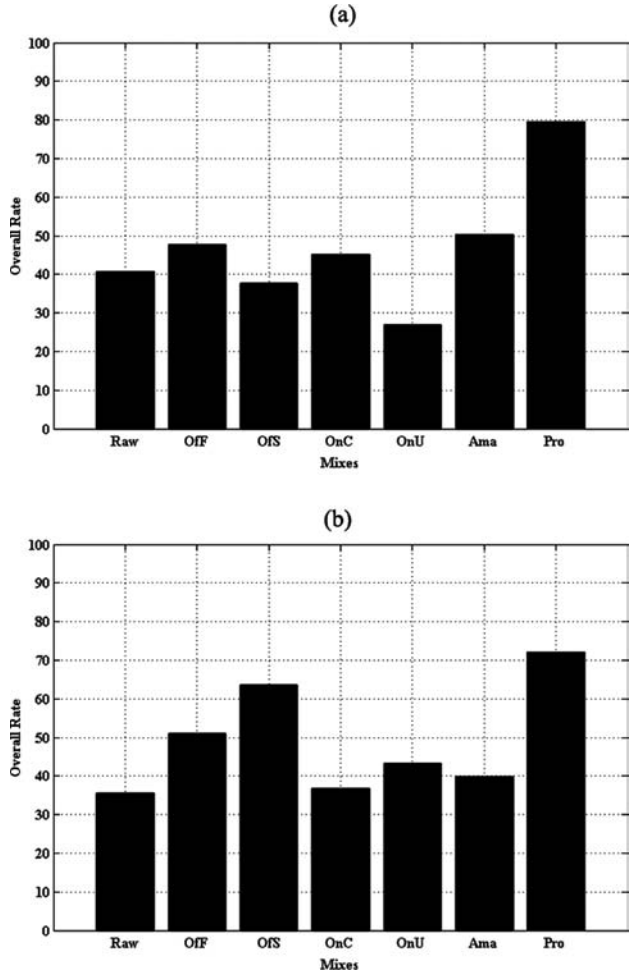

Fig. 9. Overall rating of mixes in subjective evaluation. (a) Ability to distinguish the sources (b) Overall Preference.

of attenuation to 6 dB and setting a high Q of 5. These limitations caused "Real-time Constrained" to make only minor changes to the input. In all songs, our "Offline Semi" with user controller is rated higher than "Offline Fully," which was expected since the Strength control in "Offline Semi" gives the user the chance to improve the mix made by "Offline Fully."

From Fig. 8(b), it can be seen that the overall quality rating of the "Offline Fully" and "Amateur" mixes are highly dependent on the song.

6.3 Summary

Fig. 9(a) illustrates the overall rating of each mix for Task 1. The overall ratings are achieved by averaging the ratings among all participants and songs of Task 1 for each mix. Although none of our versions has higher rating than the two manual mixes, the "Offline Fully" and "Real-time Constrained" have reduced the perceptual masking of the input multitracks as they are rated higher than "Raw" mix. The results do not illustrate an improvement by "Offline Semi" even though it was controlled by the user. The user of "Offline Semi" may have used the controller partly to improve the overall quality instead of reducing the masking. Also, perception of masking varies for individuals and some

participants may find a mix highly masked even if the mix was equalized with the purpose of masking reduction.

We also see the failure of "Real-time Unconstrained" in masking reduction. This cannot be due to the real-time nature of the system since "Real-time Constrained" is also real-time and has been successful in reducing the perceptual masking. Therefore we consider the lack of limitation on filtering and having high Q for filters as the likely reasons for the failure of "Real-time Unconstrained."

Fig. 9(b) illustrates the overall rating of each mix for Task 2, achieved by averaging the ratings among all participants and all songs of Task 2 for each mix. The results show the success of all four implementations in improving the overall quality of the mix since they are all rated higher than "Raw."

Also the offline implementations show a better performance and improvement on overall quality compared to the real-time versions. This may be due to the non-causality and/or time-invariant equalization employed by offline systems. "Offline Semi" is close to "Professional" and noticeably higher than "Offline Fully," which demonstrates the positive effect of the user parameter in "Offline Semi" for improving the overall quality. Although the real-time "Real-time Constrained" with restriction on equalization does not make a noticeable quality improvement on the input, the other real-time version without restriction on equalization ("Real-time Unconstrained") is still reliable in improving the overall quality of the input since it is rated higher than "Raw" and slightly higher than "Amateur" mix.

7 CONCLUSIONS

This paper described the automation of equalization in order to improve the overall quality by reducing the masking in a multitrack mix. Masking reduction is known to play an important role in achieving a good sounding mix. Thus we decided to test and evaluate our implementations not only based on masking reduction but also in terms of the overall preference.

Both the subjective and objective evaluations showed small changes in the amount of masking for each implementation, although the relative performance of the proposed techniques differs when assessed with objective or subjective measures. Both subjective and objective evaluations confirm the success of "Offline Fully" and "Real-time Constrained" in reducing the masking. Unexpectedly, subjective evaluation does not confirm the reduction of masking for "Offline Semi" and "Real-time Unconstrained."

We also sought to make sure that our implementations satisfy the listener in terms of overall quality, since a "good" sounding masked mix is preferred over a "bad" sounding unmasked mix. Fig. 9(b) showed that the fully automated "Offline Fully," which successfully reduced the perceptual masking both in subjective and objective evaluation, produces a mix with not only a better perceived quality than "Raw" but also a higher quality mix than an "Amateur" mix.

Comparing the average masking reduction results from Table 3 with the results from Fig. 9(b), for offline implementations, we see that small changes in the amount of masking based on the *MUR* model result in noticeable changes in the overall quality of the mix. For real-time implementation, more masking reduction based on the *MUR* model does not always lead to a more preferred mix in terms of overall quality.

Although our "Real-time Unconstrained" approach failed to reduce the perceptual masking in subjective evaluation, it resulted in the reduction of masking in objective evaluation, improvement of overall quality, and a slight preference over the "Amateur" mix.

Most importantly, we have been successful in the development of a semi-autonomous equalizer ("Offline Semi") that can be controlled by an amateur person, reduces masking according to an objective measure, and produces a mix close in overall preference to a professional mix. In other words, our semi-autonomous offline implementation has successfully simplified many complex EQ parameters per track into only one simple parameter for the whole multitrack.

It is clear that the implementations of multitrack equalization have scope for improvement. The real-time versions used filter banks, as opposed to frequency domain analysis, and thus had limited frequency resolution in the analysis stage. Though the intention of these implementations was to provide an autonomous approach to masking reduction similar to manual approaches described in the literature, it is likely that their performance could have been improved by incorporating additional knowledge from psychoacoustics. Furthermore, the approach to masking reduction is based on manual approaches and, hence, may not be considered optimal.

Further examination of the objective and subjective masking results suggest that the objective measure may be ill-suited for measuring masking in multitrack musical audio, since for all mixes of all songs, the average *MUR* value only deviated slightly from the original "Raw" average *MUR* value. Establishing a measure of masking suitable for multitrack audio production, aligned both with psychoacoustics and sound engineering practice, is clearly an important area of further research.

There is certainly scope for more rigorous and more formal subjective evaluation. Several parameters in our implementations were set based on informal listening. Rigorous method of adjustment tests would allow fine-tuning of these parameter settings to optimal values aligned with listener preference. Our tests were performed over headphones. It is well-known that headphone monitoring is an increasingly common practice [30] and often offers advantages due to avoiding influences from background noise and room acoustics [31]. However, most mixing engineers still prefer to monitor over loudspeakers, and some listening tests have shown significant differences in results between playback over headphones and over loudspeakers [32]. Thorough investigation of performance, both in terms of preference and especially masking reduction, should consider both playback methods. Finally, perceptual audio evaluation in

the form of a multistimulus test is particularly challenging when there is no clearly defined reference or low anchor. Our approach of ensuring that the whole of the scale is used addresses this problem and has been employed and discussed previously (see for instance, [11, 33]), but may also artificially inflate the importance of differences in ratings. In particular, the marginal results found in the objective testing might also be identifiable in alternative approaches to subjective evaluation.

In the process of multitrack equalization, finding the problem, identifying the problematic tracks and the spectral locations of problematic frequencies, plus choosing the proper tool for the treatment are challenging and time consuming tasks for amateurs and professional mix engineers. We have been able to successfully automate these steps, while still leaving the creative aspects of mixing to the user.

ACKNOWLEDGMENT

The authors would like to thank all volunteers from Queen Mary, University of London and elsewhere who participated in the listening tests as well as the amateur and professional mix engineers who did the manual equalizations. Thanks also to Zheng Ma for providing advice and source code for calculation of the Masked to Unmasked Ratio. This work was supported in part by EPSRC Grant EP/K007491/1, Multisource audio-visual production from user generated content.

8 REFERENCES

[1] B. Moore, "Masking in the Human Auditory System," in *Collected Papers on Digital Audio Bit Reduction*, N. Gilchrist and C. Grewin, Eds. (Audio Engineering Society, 1995).

[2] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*, 4th ed. (New York, Marcel Dekker, 2004).

[3] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed. (Springer, 2007).

[4] A. U. Case, *Sound FX: Unlocking the Creative Potential of Recording Studio Effects*, 1st ed. (Focal Press, 2007).

[5] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools* (Focal Press, 2008).

[6] M. Senior, *Mixing Secrets for the Small Studio* (Focal Press, 2011).

[7] E. Perez Gonzalez and J. D. Reiss, "Improved Control for Selective Minimization of Masking Using Inter-Channel Dependency Effects," in *11th Int. Conference on Digital Audio Effects (DAFx)*, Espoo, Finland (2008).

[8] A. Tsilfidis, et al., "Hierarchical Perceptual Mixing," presented at the *126th Convention of the Audio Engineering Society* (2009 May), convention paper 7789.

[9] B. Owsinski, *The Mixing Engineer's Handbook*, 3rd ed. (Thomson Course Technology, 1999).

[10] D. Ward, et al., "Multitrack Mixing Using a Model of Loudness and Partial Loudness," presented at the *133rd Convention of the Audio Engineering Society* (2012 Oct.), convention paper 8693.

[11] S. Mansbridge, et al., "Implementation and Evaluation of Autonomous Multitrack Fader Control," presented at the *132nd Convention of the Audio Engineering Society* (2012 Apr.), convention paper 8588.

[12] E. Perez Gonzalez and J. D. Reiss, "Automatic Gain and Fader Control For Live Mixing," presented at *the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York (2009).

[13] E. Perez Gonzalez and J. D. Reiss, "A Real-Time Semiautonomous Audio Panning System for Music Mixing," special issue on Digital Audio Effects - EURASIP Journal on Advances in Signal Processing (2010).

[14] S. Mansbridge, et al., "An Autonomous System for Multitrack Stereo Pan Positioning," presented at the *133rd Convention of the Audio Engineering Society* (2012 Oct.), convention paper 8736.

[15] P. D. Pestana and J. D. Reiss, "A Cross-Adaptive Dynamic Spectral Panning Technique," presented at the *17th Int. Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany (2014).

[16] E. Perez Gonzalez and J. D. Reiss, "Automatic Equalization of Multichannel Audio Using Cross-Adaptive Methods," presented at the *127th Convention of the Audio Engineering Society* (2009 Oct.), convention paper 7830.

[17] P. D. Pestana and J. D. Reiss, "Intelligent Audio Production Strategies Informed by Best Practices," presented at the *AES 53rd International Conference on Semantic Audio* (2014 Jan.), conference paper S2-2.

[18] V. Verfaille, et al., "Adaptive Digital Audio Effects (A-DAFx): A New Class of Sound Transformations," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1817–1831 (2006).

[19] P. Aichinger, et al., "Describing the Transparency of Mixdowns: The Masked-to-Unmasked Ratio," presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8344.

[20] S. V. Lopez and J. Janer, "Quantifying Masking in Multitrack Recordings," in Sound and Music Computing (2010).

[21] Z. Ma, et al., "Partial Loudness in Multitrack Mixing," presented at the *AES 53rd International Conference on Semantic Audio* (2014 Jan.) conference paper S2-3.

[22] ISO, "ISO 266, Acoustics–Preferred Frequencies for Measurements" (1975).

[23] B. R. Glasberg and B. C. J. Moore, "Development and Evaluation of a Model for Predicting the Audibility of Time-Varying Sounds in the Presence of Background Sounds," *J. Audio Eng. Soc.*, vol. 53, pp. 906–918 (2005 Oct.).

[24] B. R. Glasberg and B. C. J. Moore, "A Model of Loudness Applicable to Time-Varying Sounds," *J. Audio Eng. Soc.*, vol. 50, pp. 331–342 (2002 May).

[25] B. C. J. Moore, et al., "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness," *J. Audio Eng. Soc.*, vol. 45, pp. 224–240 (1997 Apr.).

[26] A. J. R. Simpson, et al., "A Practical Step-by-Step Guide to the Time Varying Loudness Model of Moore, Glasberg and Baer (1997; 2002)," presented at the *134th Convention of the Audio Engineering Society* (2013 May), convention paper 8873.

[27] B. De Man, et al.,"The Open Multitrack Testbed," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), eBrief 165.

[28] ITU, "ITU-R Recommendation BS.1534-1, Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)," International Telecommunications Union, Geneva (2001).

[29] ITU, "ITU-R Recommendation BS.1770-2, Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level," International Telecommunication Union (2011).

[30] B. Leonard, et al., "The Effect of Playback System on Reverberation Level Preference," presented at the *134th Convention of the Audio Engineering Society* (2013 May), convention paper 8886.

[31] R. King, et al., "The Effects of Monitoring Systems on Balance Preference: A Comparative Study of Mixing on Headphones versus Loudspeakers," presented at the *131st Convention of the Audio Engineering Society* (2011 Oct.), convention paper 8566.

[32] R. King, et al., "Loudspeakers and Headphones: The Effects of Playback Systems on Listening Test Subjects," in *International Congress on Acoustics (ICA)*, Montreal (2013).

[33] M. Zaunschirm, et al.,"A Sub-Band Approach to Musical Transient Modification," *Computer Music J.*, vol. 36, no. 2, pp. 23–36 (Summer 2012).

APPENDIX. PSEUDOCODE FOR REAL-TIME IMPLEMENTATION

Algorithm A.1 processBlock

```
nF = number of filters in the equalizer

// Features Extraction
for track = 1 to TotalTracks
   stereo to mono conversion
   if (Frame not silence)
      call getMagRes
      call getRank

// Masking Detection
   for masker = 1 to TotalTracks
      for maskee = 1 to TotalTracks except masker
         for rank = 1 to R_T
            if (masking occurs as defined in (1))
               update masking storage database
```

Algorithm A.1 Continued

```
// Masking Selection
for masker = 1 to TotalTracks
   call selectMasking
   for maskingIndex = 1 to nF
      if (Masking ~= 0)
         smoothen frequency and amount of masking using
          EMA
         update and store the smoothed masking frequencies and
          amounts

// Filtering
for track = 1 to TotalTracks
   for filterIndex = 1 to nF
      update filter parameters and apply filtering
// Mixing Down
Sum all input channels to the output
```

Algorithm A.2 getMagRes

```
nAF = number of analysis filters
for filterIndex=1to nAF
   Copy samples into a temporary buffer
   Filter temporary buffer
   Get RMS of filter output
   If (RMS ~= 0)
      Update Magnitude Response Database with RMS in dB
   Else
      Update Magnitude Response Database with
'close-to-zero' dB (-Inf)
```

Algorithm A.3 getRank

```
Copy Magnitude Response Database for specified track into a
temporary vector
for rank=1 to R_T
   Find and store the bin with highest magnitude in temporary
vector
   Set the magnitude of recently found bin to 'close-to-zero'
dB (-Inf)
```

Algorithm A.4 selectMasking

```
nF= number of filters in equalizer
Copy Masking Database for specified track into a temporary
vector
for maskingIndex=1 to nF
   Find and store the bin with highest magnitude in temporary
vector
   Set the masking amount of recently found bin to zero
Sort and output the selected maskings from lowest frequency
to highest
```

THE AUTHORS

Sina Hafezi

Joshua D. Reiss

Sina Hafezi received the B.Eng. degree in computer engineering in 2012 and the M.Sc. in digital signal processing in 2013 both from Queen MaryUniversity of London. He has worked in Centre for Digital Music as a researcher and software engineer on multiple projects related to autonomous equalization, which led to patent and commercial application. He started the Ph.D. in acoustic signal processing at Imperial College London in 2014.

•

Joshua D. Reiss, Ph.D., is a Reader in Audio Engineering with the Centre for Digital Music in the School of Electronic Engineering and Computer Science at Queen Mary University of London. He has bachelor's degrees in both physics and mathematics and earned his Ph.D. in physics from the Georgia Institute of Technology. He is a member of the Board of Governors of the Audio Engineering Society and co-founder of the company MixGenius. Dr. Reiss has published more than 100 scientific papers and serves on several steering and technical committees. He has investigated sound synthesis, time scaling and pitch shifting techniques, polyphonic music transcription, loudspeaker design, automatic mixing for live sound, and digital audio effects. His primary focus of research, which ties together many of the above topics, is on the use of state-of-the-art signal processing techniques for professional sound engineering.