REAL-TIME PSYCHOACOUSTIC FREQUENCY MASKING COMPENSATION FOR AUDIO SIGNALS WITH OVERLAPPING SPECTRA

Giorgio PRESTI (giorgio.presti@unimi.it) (0000-0001-7643-9915) [1],
Nicola DEGIORGI (nicoladegiorg@gmail.com) [2], **Amedeo FRESIA** (amedeofresia@gmail.com) [2], and
Antonio SERVETTI (antonio.servetti@polito.it) (0000-0002-0159-5718) [2]

[1] *Laboratory of Music Informatics (LIM)*, *Department of Computer Science*, **University of Milan**, Italy
[2] *Control and Computer Engineering Department*, **Politecnico di Torino**, Italy

ABSTRACT

This work presents a prototype for the automatic and real-time psychoacoustic frequency masking effect compensation. This tool is intended to handle audio signals with overlapping spectra coming from two different mono or stereo sources in order to emphasise those frequencies of a main source that the second one masks. The goal is achieved using a dynamic equaliser controlled by a function of the differences of the input spectra psychoacoustic models. The tool has been prototyped as a standard Virtual Studio Technology (VST3) plugin and its effectiveness has been tested with a user study carried out in the context of music production.

1. INTRODUCTION

Auditory masking is a well-known psychoacoustic phenomenon occurring when the presence of a loud sound compromises the perception of a fainter, adjacent sound by masking its presence. If the sounds are adjacent in terms of frequency components, the phenomenon is also called spectral masking, while if the sounds are temporally adjacent, it is called temporal masking [1–5]. In both cases, the amount of masking is a function of the distance between the two stimuli. By defining the distance measure and the masking function, different psychoacoustic models can be implemented, varying the degree of approximation according to the desired goals.

An example of a common psychoacoustic model is that used in the MPEG-2 audio compression [6]; a lossy compression technique which exploits a non-uniform quantization of the frequency content of a signal such that the quantisation noise remains under a certain Masking Threshold (MT), thus reducing the amount the information without affecting the perceived quality. The MP3 psychoacoustic model is devoted to the computation of the MT and can operate at different degrees of approximation, allowing a trade-off between encoding time and quality (not to be confused with bitrate).

In particular, the frequency masking contribution of a psychoacoustic model can account for a lot of different factors (loudness, frequency register, spectral complexity, and so on), but in its most simple definition, it says that the loudest frequency inside a critical band [7, 8] masks the fainter ones inside the same band. This phenomenon can be used to approximate an MT by convolution [9] or matrix multiplication [10] of the spectrum in the Bark scale [11] with a set of Spreading Functions (SF) – which is an asymmetrical triangular-like function the size of a critical band.

The goal of this work is to exploit such a mechanism to emphasize the masked frequencies of a signal in a real-time context, where the masker and the masked sounds are available as separate streams, such as tracks inside a Digital Audio Workstation (DAW). In the context of music production, this will allow the sound engineer to resolve the situation where an accompanying instrument masks a lead instrument [12], *e.g.* a guitar masking a vocal track.

To achieve this goal, a VST3 plugin called TheMasker has been implemented, taking as input the signal that will be processed and as a side-chain the signal that is potentially masking the input. An MT is computed from the side-chain, and then a dynamic equalizer boosts the input components lower than the MT. The user is provided with enough controls to fine-tune the process, but to avoid information overload, many algorithm parameters are not exposed, instead, they have been tuned empirically to adapt to most situations.

The novelty of the proposed approach lies in the use of a psychoacoustic model together with the possibility of managing the masking signal as a simple side chain input, which is a combination not sufficiently explored in the literature, as discussed in Section 2.

The context of music production has been chosen as a reference in order to ease testing (*i.e.* giving the plugin to professionals and collecting their feedback), but also other contexts may benefit from such processing, examples will be discussed in the conclusions.

The remainder of the paper is organized as follows: first, existing solutions will be examined in Section 2, then the implementation of TheMasker is described in Section 3, an evaluation campaign is reported and discussed in Section 4, and finally Section 5 concludes the paper with additional remarks and future works.

2. EXISTING SOLUTIONS

The problem tackled in this context has already been discussed in the scientific literature, and also commercial products are already available on the market.

In general, a naive approach to the problem would be to use dynamic equalisers or multiband dynamics processors with the masker signal in the side chain to expand the portion of the masked signal whenever the masker is above a certain threshold. This threshold can be set manually (a general-purpose tool can be used in this sense) or automatically, such as in the work of Sack *et al.* [13], or in *MAuto-DynamicEq* [14]. Nevertheless, none of these approaches is based on a proper psychoacoustic model.

Finer processing may be achieved by working with a non-parametric representation of the target equalisation curve, such as in *TrackSpacer* [15], that uses a 32-band, fixed-frequency dynamic equaliser to apply an equalisation which corresponds to the flipped spectrum of the side chain input, but again, no explicit psychoacoustic model is mentioned.

Also Ahmetovic *et al.* [16] uses a non-parametric spectral representation, with a perceptually-motivated band choice in order to apply a gain to each band such that a target signal-to-noise ratio is reached.

A famous hardware processor that indeed accounts for psychoacoustic effects is the Vitalizer [17] (also available as VST plugin), which uses a patented model to achieve (among other things) a de-masking of those frequencies masked by other components of the same signal. Unfortunately, no side-chain is contemplated in this tool.

Gonzalez and Reiss [18] and Hafezi and Reiss [19] have proposed processors that use an elegant psychoacoustic model of interacting tracks; the first operating a broadband gain change of a target track, while the second implementing a sophisticated modulation of the parameters of a set of 5-band parametric equalisers on each examined track, minimizing their masking interactions. Note that this "multitrack" approach is more relevant in the context of automatic mixing, while constraining the intervention on a single track may be more suited for the realisation of a traditional mixing workflow tool.

The proposed solution is based on a simpler psychoacoustic model to that proposed by Gonzalez, Hafezi, and Reiss. Still, it uses a correction curve described by a 32-band filterbank, allowing finer control of the frequency content. Moreover, a two-input single-output routing scheme allows the integration in all those DAWs that do not support the multiple-input multiple-output routing required by Hafezi and Reiss.

3. IMPLEMENTATION

TheMasker has been first prototyped in Matlab and then implemented in C++ using the JUCE framework. In Figure 1 a block diagram of the signal processing strategy is visible: the side-chain signal is used to compute a frequency domain MT, and then the input spectrum is compared against the MT in a per-band logic. The difference in decibels of each band is used to control the gains of a Linkwitz-Riley band-pass filterbank according to the settings decided by the user.

Ideally, both the side-chain send and the plugin position in the channel processing chain should be post-fader, so to consider the actual mixing level, thus obtaining a realistic effect, nevertheless, the user is left with the ability to tweak side-chain, input, and output levels, exposed as plugin parameters.

A Graphical User Interface (GUI), visible in Fig. 2, has also been implemented to ease the testing campaign of the realised tool.

3.1 Masking Threshold Estimator

The MT is computed by taking a Flat-Top-windowed 1024-point 50%-overlap Fast Fourier Transform (FFT), resulting in 512 magnitude values of the side-chain input. Then, the resulting signal frequencies are spaced in the Bark scale by matrix multiplication with a 32×512 filterbank matrix, and converted in the Decibel scale. A reduction to a 32 values array is finally realised by matrix multiplication with a 32×32 SF matrix, providing a 32-points approximation of the side-chain MT.

The used SFs are expressed in dB and are centred on 32 linearly-spaced Bark frequencies f_b, with a rising slope of $+27$dB/Bark and a falling slope of -12dB/Bark [10, 20] that approximate the following SF expression from Painter and Spanias [21, 22]:

$$\mathrm{SF}(f_b) = 15.81 + 7.5(f_b + 0.474) - 17.5\sqrt{1 + (f_b + 0.474)^2} \tag{1}$$

an example of Eq. 1 is visible in Figure 3.

3.2 Input Correction Estimator

Similarly to the first stages of side-chain processing, the main input signal is taken to the 32-points Bark/dB space, but without any SF applied to it. This input representation is compared to the MT by means of a simple subtraction: positive delta values indicate that that band is masked by the side-chain, while negative delta values mean the input is not masked.

At this point, positive delta values should be directly used as gain dB values to ensure the masked input bands reach the MT, nevertheless, many issues may arise by doing so, therefore a number of constraints and corrections are enforced.

Silent or very quiet parts of the input spectrum would be amplified by a disproportionate amount than expected if in coincidence with a non-quiet part of the side-chain spectrum. This issue has been avoided by multiplying the delta values with an Intervention Limiting Function (ILF), computed as a function of the input representation. In particular, the ILF is computed as:

$$\mathrm{ILF}(n) := 0.5 \left(1 + \tanh\left(\frac{x(n) - \tau}{k}\right)\right) \tag{2}$$

where $x(n)$ is the n-th band of the input representation, τ is a threshold value and k is a knee parameter. This function scales the delta values to 0 when the input is lower than τ,

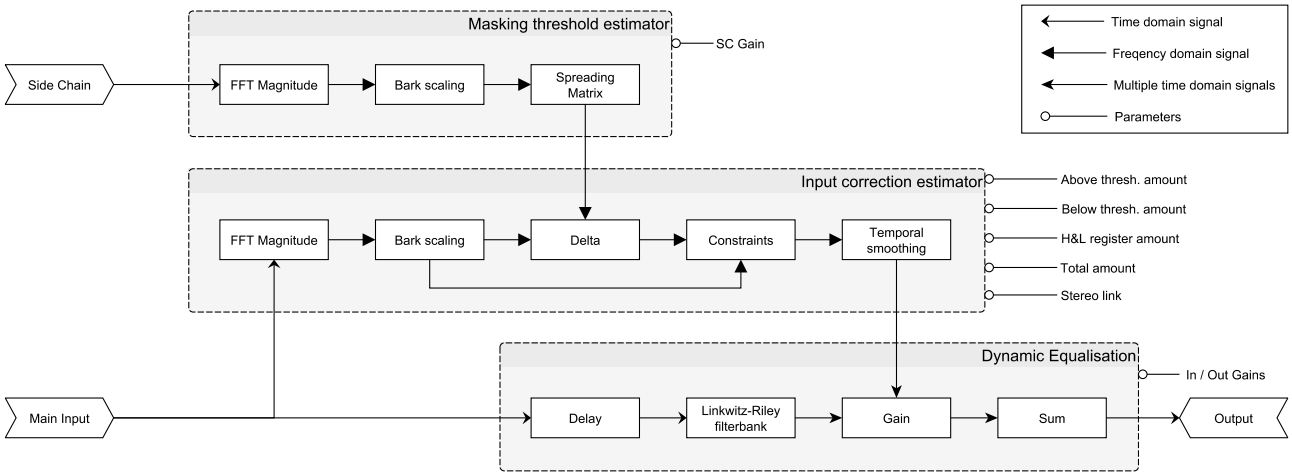

Figure 1. TheMasker block diagram.

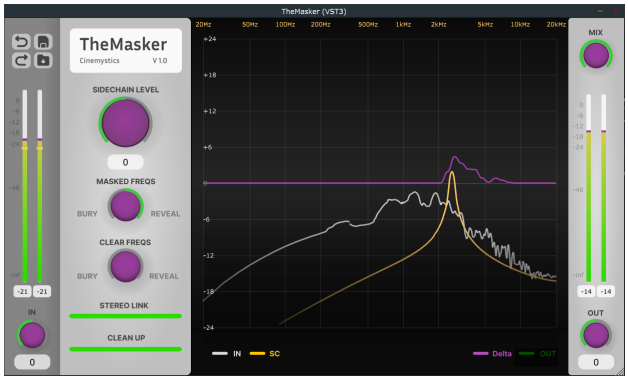

Figure 2. TheMasker Graphical User Interface. In this example, a sinewave masking some noise is visible, with the correction curve boosting all noise components below a spreading window.

and scales the delta values to their original value when the input is above τ. This behaviour is smoothed according to k, which ensures a gradual transition between the gated and the non-gated part. A τ of -40db with a $k = 3.2$ has been chosen since it produces a smooth transition between -50dB and -30dB, which has been empirically found to be adequate for most situations.

The same issue is actually present also in the case of the side-chain being very quiet, thus a second ILF with the same τ and k parameters, but computed on the side-chain instead of the input signal, is applied to the delta values.

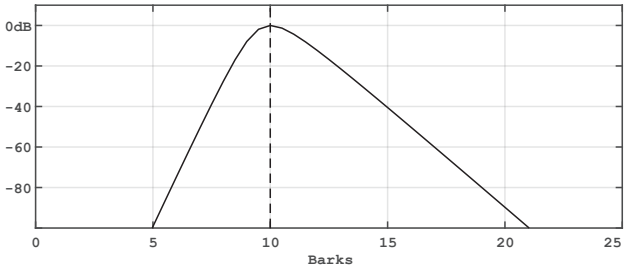

Figure 3. A spreading function centered on 10 Barks, as defined in Eq. 1

A second similar issue may arise even if the masked frequencies would genuinely benefit from a boost, but the amount of gain is so high that it may denature the sound in a perceivable way. For this reason, the delta values are soft-clipped between $\pm$12dB, again by means of an *tanh* function ($\delta(n)$ is the n-th band delta value):

$$\delta(n) = 12 \cdot \tanh\left(\frac{\delta(n)}{12}\right) \tag{3}$$

Finally, the delta values are converted to a correction equalisation curve by letting the user decide the scaling of the intervention in a number of ways:

- A scaling factor for the positive delta values, allowing the demasking of the input;
- A scaling factor for the negative delta values, allowing for the attenuation for the non-masked input components;
- A scaling factor for the overall delta values, to tune the overall intervention (exposed as an "amt" parameter);
- A frequency-dependent scaling factor allowing the user to process the whole spectrum or excluding the low and high register from being processed (exposed as a "Cleanup" parameter);

An additional "Stereo link" parameter allows the user to process the stereo channels independently or by averaging left and right delta values.

To avoid audible artefacts occurring in case of fast changes, the output values are smoothed in the time domain with different linear ramps for attack and release phases, set respectively to 80ms and 250ms. These values have been chosen empirically to maximise intervention transparency.

3.3 Dynamic Equalisation

First, to compensate for the latency introduced by the buffering needed to compute the FFT, a delay of 1024 samples is performed on the input signal. According to the VST3 standard, this delay is declared to the host application to let the DAW compensate for the introduced latency.

To achieve the dynamic equalisation the input signal is split into 32 bands via a Linkwitz-Riley crossover filterbank, with each band centred on the corresponding Bark scale representation frequency. Each band is then multiplied with the smoothed values provided by the Input Correction Estimator described in Section 3.2, compensating the cumulative effects of neighbour bands as described in [23], and then summed back to a single signal.

4. TESTING

4.1 Protocol

To test the effectiveness of TheMasker, a set of 6 stimuli has been administered to a pool of 13 subjects together with a questionnaire regarding intelligibility and perceived quality.

In particular, each stimulus is composed of three versions of two mixes. The two mixes are composed of the following signals:

- Masker: distorted guitar; masked: voice.
- Masker: long-term average speech spectrum (LTASS) noise [24]; masked: voice.

The first stimulus has been chosen as an ecological use case in the context of music production, while the latter is a synthetic example of speech enhancement in a noisy environment context.

The baseline version is an unprocessed mix of the tracks set to have the masked signal 3dB lower than the masker in terms of perceived loudness (the masker level is set to be -28 LUfs); in the second version the tracks are set to the same level of the baseline, but the masked signal is processed with TheMasker to let the masked frequencies reach the same level of the corresponding masker bands; finally, the last version is the same mix of the baseline but with a linear gain of $+3$dB on the masked signal.

The three versions of a single mix are presented together through closed headphones, with the possibility to switch between each version. For each mix, the subjects have to evaluate through a 5 points Likert scale the three versions according to the following questions:

1. How would you rate intelligibility?
2. Does the voice sound natural (1) or processed (5)?
3. How similar are the mixing levels if compared to this reference track?

For the final question, the user is given a new "reference" track which is a copy of the baseline. In this way, it is possible to check the reliability of a subject answers based on her ability to identify the baseline as identical to the reference.

The ideal outcome consists of TheMasker scores revealing an intelligibility improvement greater or equal to that of the simple volume increase correction, without too much loss in naturalness. Moreover, the version processed with TheMasker is expected to be more similar to the baseline in terms of mixing level similarity.

4.2 Results

First, it has been checked if some subjects failed to recognise the baseline and the reference as the same mix. It emerged that 3 subjects failed this task by giving a score ≤ 3 to the mix similarity of the baselines. These subjects were removed from the analysis, thus reducing the number of available subjects to 10 units.

Then a Kruskal-Wallis test was run for all 3 measures (intelligibility, naturalness and mix similarity) to see if answers relative to the 3 versions of the mixes (labelled Baseline, TheMasker, and Level in Fig. 4) are significantly different.

From the Kruskal-Wallis test, it emerged that indeed, intelligibility and mix similarity have been scored differently for the three versions ($p < 0.001$), while naturalness presents no statistically significant differences ($p = 0.087$).

Concerning intelligibility, Baseline is significantly less intelligible than the other two versions ($p < 0.005$) having a median score of 2, versus a median score of TheMasker and Level of 4 (TheMasker and Level have no significant differences, with $p = 0.768$)

The unexpected absence of significant differences in the perceived naturalness of sound (especially between Baseline and the other versions) can be interpreted as a kind of processing respectful of the original sound, which is understandable for the Level, but surprising for TheMasker.

Finally, the perception of mix similarity of TheMasker (median $= 3.5$) and Level (median $= 2$) resulted to be different ($p < 0.05$) in favour of the TheMasker, being evaluated more similar to the references than the Level version.

From these results, it can be deduced that TheMasker introduces intelligibility improvements comparable to those of a level increase, with the same impact on perceived timbre, but with less obvious changes in mix balance, thus demonstrating the usefulness of the proposed tool.

5. CONCLUSIONS

A prototype for the real-time compensation of psychoacoustic frequency masking has been implemented and tested in a real-world music-production scenario with promising results. The prototype is available for download at https://www.lim.di.unimi.it/demo_vst_eng.phpfor macOS, Windows, and Linux platforms.

Future works will focus on the improvement of the psychoacoustic model and the quality of the dynamic equalisation engine.

Also, other use cases of the proposed prototype will be

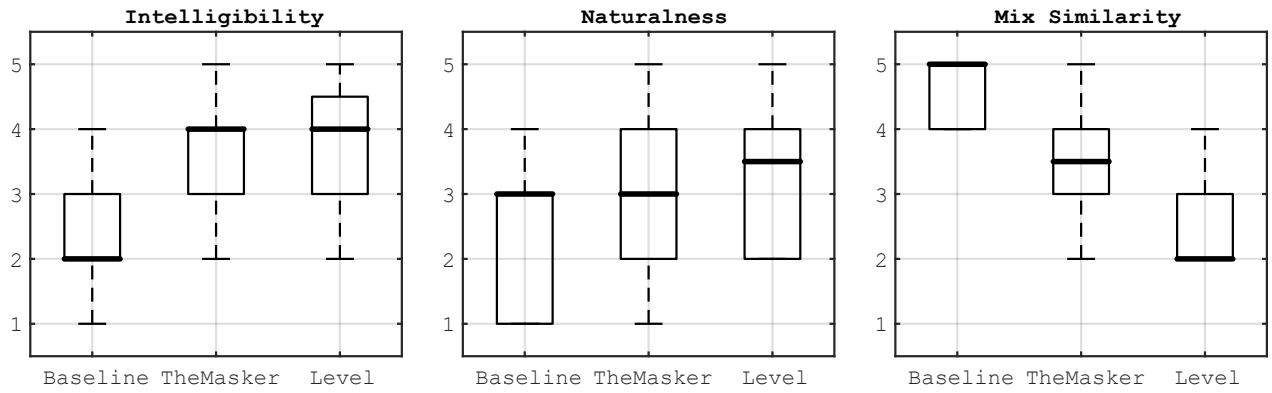

Figure 4. Results of the validation campaign: answers of 10 subjects when asked to compare a Baseline mix with two enhanced versions (one with a vocal track processed with TheMasker and one where the vocal track has only been amplified).

explored in the future, such as real-time intelligibility improvements of headphone output in noisy scenarios.

6. REFERENCES

[1] B. C. J. Moore, "Masking in the human auditory system," *Journal of the Audio Engineering Society*, May 1996.

[2] V. Best, "An introduction to psychological and physiological acoustics," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. A133–A134, 2021.

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

[4] A. J. Oxenham, "Mechanisms and mechanics of auditory masking," *The Journal of physiology*, vol. 591, no. Pt 10, p. 2375, 2013.

[5] J. P. Egan and H. W. Hake, "On the masking pattern of a simple auditory stimulus," *The Journal of the Acoustical Society of America*, vol. 22, no. 5, pp. 622–630, 1950.

[6] I. J. S. 29, "Generic coding of moving pictures and associated audio information — part 3: Audio," Genève, Switzerland, Tech. Rep., 1998.

[7] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *Bell System Technical Journal*, vol. 12, no. 4, pp. 377–430, 1933.

[8] H. Fletcher, "Auditory patterns," *Reviews of modern physics*, vol. 12, no. 1, p. 47, 1940.

[9] J. Herre and S. Dick, "Psychoacoustic models for perceptual audio coding—a tutorial review," *Applied Sciences*, vol. 9, no. 14, p. 2854, 2019.

[10] G. Schuller, *Psycho-Acoustic Models*. Springer International Publishing, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-51249-1_4

[11] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

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

[13] M. C. Sack, S. Buchinger, W. Robitza, P. Hummelbrunner, M. Nezveda, and H. Hlavacs, "Loudness and auditory masking compensation for mobile tv," in *2010 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2010, pp. 1–6.

[14] MeldaProduction. (2020) Mautodynamiceq. [Online]. Available: https://www.meldaproduction.com/MAutoDynamicEq

[15] Wavesfactory. (2014) Trackspacer. [Online]. Available: https://www.wavesfactory.com/audio-plugins/trackspacer/

[16] D. Ahmetovic, G. Galimberti, F. Avanzini, C. Bernareggi, L. A. Ludovico, G. Presti, G. Vasco, and S. Mascetti, "Enhancing screen reader intelligibility in noisy environments," *IEEE Transactions on Human-Machine Systems*, 2023.

[17] S. electronics. (2000) Vitalizer. [Online]. Available: https://spl.audio/en/spl-produkt/stereo-vitalizer-mk2/

[18] E. P. Gonzalez, J. D. Reiss *et al.*, "Improved control for selective minimization of masking using interchannel dependancy effects," in *11th Int. Conference on Digital Audio Effects (DAFx)*, 2008, p. 12.

[19] 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, May 2015.

[20] G. Schuller. (2016) Psychoacoustics models. [Online]. Available: https://colab.research.google.com/github/GuitarsAI/AudioCodingTutorials/blob/master/AC_05_psychoAcousticsModels.ipynb

[21] T. Painter and A. Spanias, "A review of algorithms for perceptual coding of digital audio signals," in *Proceedings of 13th International Conference on Digital Signal Processing*, vol. 1, 1997, pp. 179–208 vol.1.

[22] 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.

[23] V. Välimäki and J. D. Reiss, "All about audio equalization: Solutions and frontiers," *Applied Sciences*, vol. 6, no. 5, p. 129, 2016.

[24] A. Löfqvist, "The long-time-average spectrum as a tool in voice research," *Journal of phonetics*, vol. 14, no. 3-4, pp. 471–475, 1986.