INTELLIGENT SYSTEMS FOR MIXING MULTICHANNEL AUDIO

Joshua D. Reiss

Centre for Digital Music, Queen Mary University of London

ABSTRACT

Multichannel signal processing techniques are usually concerned with extracting information about sources from several received signals. In this paper, we describe an emerging field of multichannel audio signal processing where the inter-channel relationships are exploited in order to manipulate the multichannel content. Applications to real-time, automatic audio production are described and the necessary technologies and the architecture of such systems are presented. The current state of the art is reviewed, and directions of future research are also discussed.

Index Terms— Real-time, multichannel audio signal processing, automatic mixing, cross-adaptive digital audio effects.

1 INTRODUCTION

Rapid growth in the quantity of unprocessed audio material has resulted in a similar growth in the engineering tasks and requirements that must be addressed during the audio production process. Although the production tasks are challenging and technical, much of the initial work follows established rules and best practices. Yet multichannel audio content is still often manipulated 'by hand,' using no computerised signal analysis. This is a time consuming process, and prone to errors. Only if time and resources permit, does the sound engineer refine his choices to produce an aesthetically pleasing mix which best captures the intended sound.

In order to address this challenge, a new form of multichannel audio signal processing has emerged. Intelligent tools have been devised that analyse the relationships between all channels in order to automate the mixing of multichannel audio content. By 'intelligent', we mean that these are expert systems that perceive, reason, learn, and act intelligently. This implies that they must analyse the signals upon which they act, dynamically adapt to audio inputs and sound scene, automatically configure parameter settings, and exploit best practices in sound engineering to modify the signals appropriately. They derive the parameters in the editing of recordings or live audio based on analysis of the audio content, and based on objective and perceptual criteria.

For progress towards intelligent systems in these sound engineering domains, significant problems must be overcome that have not yet been tackled by the research community. Most state of the art audio signal processing techniques focus on single channel signals. Yet multichannel signals are pervasive, and the interaction and dependency between channels plays a critical role in audio production quality. This issue has been addressed in the context of audio source separation research, but the challenge in source separation is generally dependent on how the sources were mixed, not on the respective content of each source. New, multi-input multi-output audio signal processing methods are required, which can analyse the content of all sources in order to improve the quality of capturing, editing and combining multichannel audio.

2 STATE OF THE ART IN MIXING MULTICHANNEL AUDIO

The idea of automating the audio production process, although relatively unexplored, is not new. In [1], the editor of Sound on Sound magazine wrote, "There's no reason why a band recording using reasonably conventional instrumentation shouldn't be EQ'd and balanced automatically by advanced DAW software." He also wrote '[audio interfaces can] come with a 'gain learn' mode... DAWs could optimise their own mixer and plug-in gain structure while preserving the same mix balance." This would address the needs of the musician who doesn't have the time, expertise or inclination to perform all the audio engineering required. Similarly, [2] introduced the concept of an Intelligent Assistant, incorporating psychoacoustic models of loudness and audibility, intended to "take over the mundane aspects of music production, leaving the creative side to the professionals, where it belongs."

Currently, multichannel audio editing tools demand manual intervention. Although audio editors are capable of saving a set of static scenes[3] for later use, they lack the ability to take intelligent decisions, such as adapting to different acoustic environments or different set of inputs. An exception is the Intelligent Audio Editor [4], but it requires a machine-readable score, and hence is highly limited in its application.

One attempt to reduce the required effort of the sound engineer while mixing multichannel content is the

978-1-4577-0274-7/11/$26.00 ©2011 IEEE DSP2011

development of automatic riders. This is a type of gain control which continually and smoothly adjusts the gain on a channel in order to match a given criterion. Recently riders have been devised that, given the desired loudness level of a target channel in relation to the rest of the mix, will compensate for all deviations by raising or lowering the levels of the target (e.g. the Vocal Rider from Waves Ltd.). The existing riders only work with at most two channels of audio, and thus riders have limited flexibility; all channels other than the target are combined into a single channel, only the target is modified, and only a single loudness curve can be used.

Research has also focused on a mixture of a large number of audio signals where data compression is achieved by removing sounds that are masked by other sources [5-6]. Although relevant for real-time audio applications where data size is an issue, it is only concerned with mixes produced by summing the sources.

Interactive audio applications have also provided a relevant approach to automation. In game audio, the user may change his location with respect to the sources, and sources are then rendered to give the appropriate spatial characteristics [7].

2.1 Intelligent and Adaptive Digital Audio Effects

Rather than have sound engineers manually apply many audio effects to all audio inputs and determine their appropriate parameter settings, intelligent, adaptive digital audio effects may be applied instead [8]. The parameter settings of adaptive effects are determined by analysis of the audio content, where the analysis is achieved by a feature extraction component built into the effect. Intelligent audio effects also analyse or 'listen' to the audio signal, but are furthermore imbued with knowledge of their intended use, and control their own operation in a manner similar to manual operation by a trained engineer. The knowledge of their use may be derived from established best practices in sound engineering, psychoacoustic studies that provide understanding of human preference for audio editing techniques or machine learning from training data based on previous use. Thus, an intelligent audio effect may be used to set the appropriate equalisation, automate the parameters on dynamics processors, and adjust stereo recordings to more effectively distinguish the sources. A block diagram of an intelligent audio effect is given in Fig. 1.

The side chain is essential for low latency, real time signal processing flow. The audio signal flow is unaffected, but any required analysis, is performed in a separate analysis section. The side-chain is comprised of a feature extraction section and an analysis section that processes the features.

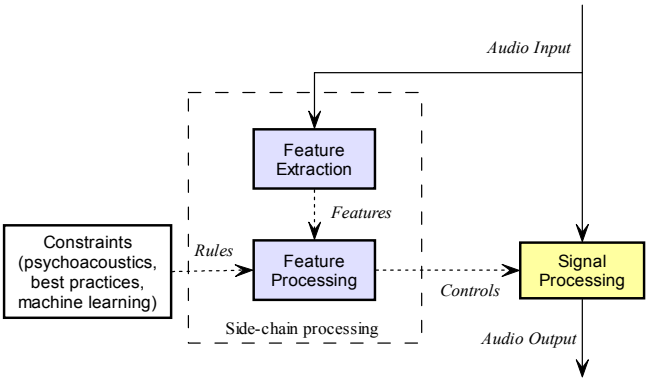

Figure 1. Block diagram of an intelligent audio effect. Features are extracted by analysis of the audio signal. These features are then processed based on a set of rules intended to mimic the behavior of a trained engineer. A set of controls are produced which are used to modify the audio signal.

The feature extraction is in charge of extracting a series of features from the input channel. Accumulative averaging, described in a later section, is used to ensure real time signal processing operations, even when the feature extraction process is non-real time. The analysis section outputs control signals to the signal processing side in order to trigger the desired parameter control change command.

Recently, we introduced several intelligent, adaptive effects, for use with single channel audio, which automate many parameters and enable a higher level of audio editing and manipulation [9-10]. This included adaptive effects that control the panning of a sound source between two user-defined points depending on the sound level or frequency content of the source, and noise gates with parameters which are automatically derived from the signal content.

2.2 Cross-Adaptive Digital Audio Effects

When editing multichannel audio, one performs signal processing changes on a given signal source not only because of the source content but also because there is a simultaneous need to blend it with the content of other sources, so that a high quality mix is achieved. The relationship between all the sources involved in the audio mix must be taken into account. Thus, a cross-adaptive effect processing architecture is ideal for automatic mixing.

In a cross-adaptive effect, also known as inter-channel dependent or MIMO (multi-input / multi-output) effect, the signal processing of an individual source is the result of the relationships between all involved sources. That is, these effects analyse the signal content of several input channels in order to produce several output channels. This generalizes the single channel adaptive signal processing mentioned above.

In an intelligent multichannel audio editing system, as shown in Fig. 2, the side-chain will consist of a feature extraction section for each channel and a single analysis

section that processes the features extracted from many channels. The cross-adaptive processing section of an intelligent multichannel audio editing system exploits the interdependence of the input features in order to output the appropriate control data. This data controls the parameters in the signal processing of the multichannel content. The cross-adaptive feature processing can be implemented by a set of constrained rules that consider the interdependence between channels.

In principle, cross-adaptive digital audio effects have been in use since the development of the microphone mixer [11]. However, such systems are only concerned with automatic gain handling and require a significant amount of human interaction during setup to ensure a stable operation.

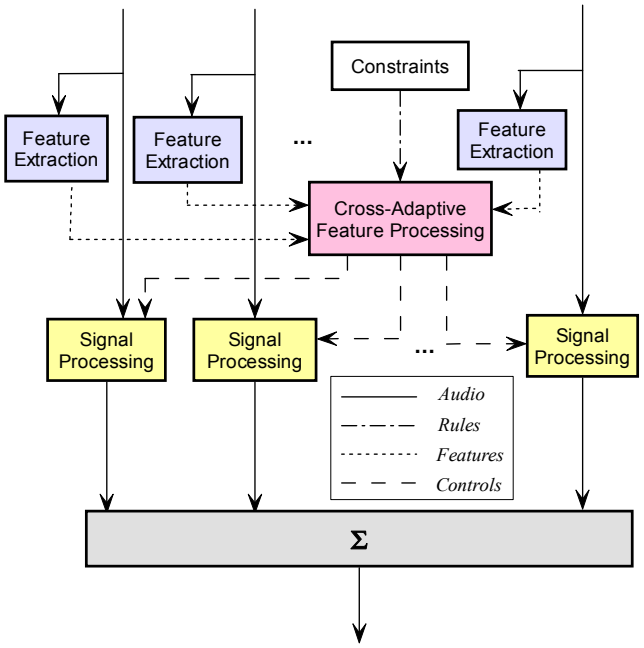

Figure 2. Block diagram of an intelligent, cross-adaptive mixing system. Extracted features from all channels are sent to the same feature processing block, where controls are produced. The output channels are summed to produce a mix that depends on the relationships between all input channels.

2.3 Intelligent, Multichannel Digital Audio Effects

In [12], and references therein, several cross-adaptive digital audio effects were described that explored the possibility of reproducing the mixing decisions of a skilled audio engineer with minimal or no human interaction. Each of these effects produces a set of mixes where each output may be given by the following equation;

$$mix_l(n) = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} c_{k,m,l}(n) * x_m(n) \,. \qquad (1)$$

That is, the resultant mixed signal at time *n* is a sum over all input channels, of a control vectors convolved with the input signal. For automatic faders and source enhancement, the control vectors are simple scalars, and hence the convolution operation becomes multiplication. For polarity correction, a binary valued scalar, ±1, is used. For automatic panners, two mixes are created, where panning is also determined with a scalar multiplication (the sine-cosine panning law). For delay correction, the control vectors become a single delay operation. This applies even when different delay estimation methods are used, or when there are multiple active sources [13]. And automatic equalization employs impulse responses for the control vectors based on transfer functions representing each equalization curve applied to each channel.

In [14], a source separation technique was described where the control vectors are impulse responses that represent IIR unmixing filters for a convolutive mix. Thus, each of the resultant output signals, $mix_l(n)$ in Eq. (1) represents a separated source, dependent on filtering of all input channels.

In fact, any cross-adaptive digital audio effect that employs linear filters may be described in this manner. Multichannel dynamic range compression would be based on a time varying gain for each control vector based on analysis of the relative loudness range of each input channel, and in multichannel reverberation the control vectors would be represented as static, finite impulse responses.

The approach taken in [15] attempted to deliver a live monitor mixing system that was as close as possible to a predefined target. It approximated the cause and effect relationship between inputs to monitor loudspeakers and intelligibility of sources at performer locations. The stage sound was modelled as a linear multi-input multi-output system which enabled all performer requirements to be considered simultaneously. Simple attenuation rules from the monitors to the performers were used in order to perform simulations in a free field environment. The target mix was defined in terms of relative sound pressure levels (SPLs) for each source at each performer. Thus, a constrained optimization approach was used to generate scalar valued control vectors that resulted in optimized mixes at selected positions on stage.

The reverse engineering of a mix, described in [16], assumes that Eq. 1 holds when presented with original tracks and a final mix. It then uses least squares approximation to estimate each control vector as a fixed length FIR filter. By assuming that the gain in the filter represents the faders, the initial zero coefficients represent delay, differences between left and right output channels are based on sine-cosine panning, and that anything remaining represents equalization, it can then reverse engineer the settings used for time-varying faders, panning, delays and equalization. However, this would not be considered an intelligent audio effect since it requires little or no knowledge of preferred mixing decisions.

3 REAL-TIME, MULTICHANNEL INTELLIGENT AUDIO SIGNAL PROCESSING

The standard approach adopted by the research community for real-time audio signal processing is to perform a direct translation of a computationally efficient off-line routine into one that operates on a window by window basis. However, effective use in live sound or interactive audio requires not only that the methods be real-time, but also that there is no perceptible latency. The minimal latency requirement is necessary because there should be no perceptible delay between when a sound is produced and when the modified sound is heard by the listener. Thus, many common real-time technologies, such as look-ahead and the use of long windows, are not possible. The windowed approach produces an inherent delay (the length of a window) that renders such techniques impractical for many applications. Nor can one assume time invariance; sources move and content changes during performance. To surmount these barriers, perceptually relevant features must be found which can be quickly extracted in the time domain, analysis must rapidly adapt to varying conditions and constraints, and effects must be produced in advance of a change in signal content.

In this section, we look at some of the main enabling technologies that are used. An excellent, more detailed review may be found in [17].

3.1 Reference Signals and Adaptive Thresholds

An important consideration to be taken into account during analysis of an audio signal is the presence of noise. The existence of interference, crosstalk and ambient noise will influence the ability to derive information about the source. For many tasks, the signal analysis should only be based on signal content when the source is active, and the presence of significant noise can make this difficult to identify.

One of the most common methods used for ensuring that an intelligent tool can operate with widely varying input data is adaptive gating, where a gating threshold adapts according to the existing noise. A reference microphone placed far from the source signal may be used to capture an estimation of ambient noise. This microphone signal can then be used to derive the adaptive threshold. Although automatic gating is typically applied to gate an audio signal it can also be used to gate whether the extracted features will be processed.

The most straightforward way to implement this is to apply a gate that ensures that the control vector is only updated when the signal level of the m^{th} channel is larger than the level of the reference, as given in the following equation;

$$c_m(n+1) = \begin{cases} c_m(n) & x_{m,RMS}{}^2(n) \le r_{RMS}{}^2(n) \\ \alpha c'_m(n+1) + (1-\alpha)c_m(n) & \text{otherwise} \end{cases} \quad (2)$$

Where c' represents an instantaneous estimation of the control vector. Thus the current control vector is a weighted sum of the previous control vector and some function of the extracted features. Initially, computation of RMS level of a signal x is given by

$$x_{RMS}{}^2(n) = \frac{1}{M}\sum_{m=0}^{M-1} x^2(n-m)\,. \quad (3)$$

And later values may either be given by a sliding window, which reduces to

$$x_{RMS}{}^2(n+1) = \frac{x^2(n+1) - x^2(n+1-M)}{M} + x_{RMS}{}^2(n)\,, \quad (4)$$

or a low-pass one pole filter (also known as an exponential moving average filter) ,

$$x_{RMS}{}^2(n+1) = \beta x^2(n+1) + (1-\beta)x_{RMS}{}^2(n)\,. \quad (5)$$

 and represent time constants of IIR filters, and allow for the control vector and RMS estimation, respectively, to smoothly change with varying conditions. Eq. (4) represents a form of dynamic real-time extraction of a feature (in this case, RMS), and Eq. (5) represents an accumulative form.

3.2 Incorporating Best Practices into Constrained Control Rules

In order to develop intelligent software tools, it is essential to formalise and analyse audio production methods and techniques. This will establish required functionality of such tools. Furthermore analysis of the mixing and mastering process will identify techniques that facilitate the mixing of multi-channel tracks, and repetitive tasks which can be automated. By establishing methodologies of audio production used by professional sound engineers, features and constraints can be specified that will enable automation.

Many of the best practices in sound engineering are well-known, and have been described in the literature [18]. In live sound for instance, the maximum acoustic gain of the lead vocalist, if present, tends to be the reference to which the rest of the channels are mixed, and this maximum acoustic gain is constrained by the level at which acoustic feedback occurs. Furthermore, resonances and background hum should be removed from individual sources before mixing, all active sources should be heard, delays should be set so as to prevent comb filtering, dynamic range compression should reduce drastic changes in loudness of one source as compared to the rest of the mix, panning should be balanced, spectral and psychoacoustic masking of sources must be minimised, and so on.

Similarly, many aspects of sound spatialisation obey standard rules. For instance, sources with similar frequency content should be placed far apart, in order to prevent spatial

masking and improve the intelligibility of content. A stereo mix should be balanced and hard panning is avoided. When spatial audio is rendered with height, low frequency sound sources are typically placed near the ground, and high frequency sources are placed above, in accordance with human auditory preference. Also, many parameters on digital audio effects can be set based on analysis of the signal content, e. g., attack and release on dynamics processors are kept short for percussive sounds.

These best practices and common approaches translate directly into constraints that are built into intelligent software tools. For example, [19] uses a measure of the spectral masking to enhance a source while minimizing the changes in levels of the other sources, similar to the way in which a sound engineer attempts to make only slight changes to those sources that make a mix sound 'muddy.' This is achieved by using filterbanks to find the dominant frequency range of each input signal. If f_C is the dominant frequency of the channel that we wish to enhance, and f_m is the dominant frequency of the m^{th} channel, where $m \neq C$, then the control vector applied to this channel is given by a Gaussian function;

$$c_m = G \frac{1}{Q\sqrt{2\pi}} e^{-(f_m(n)-f_C(n))^2/(2Q^2)} - 1, \qquad (6)$$

where dynamic or accumulative averaging, as described previously, may be used to avoid rapid variation in estimation of dominant frequencies, while still adapting to changing content.

Eq. (6) allows channels with closely related frequency content to be highly attenuated, while minimizing the attenuation of those channels with frequency content far from the source, e. g., if one wants to enhance the stand-up bass in a mix, the kick drum will be attenuated but there will be little change to the flute. This method also uses knowledge of sound engineering practice since its operation is similar to that of a parametric equalizer, commonly used for boosting or attenuating a frequency range in single channel audio processing. That is, G controls the amount of attenuation, f_C represents the centre frequency, and Q the quality.

4 FUTURE WORK

4.1 Psychoacoustic Studies

Important questions concerning the psychoacoustics of mixing multi-channel content. For instance, little has been formally established concerning user preference for relative amounts of dynamic range compression used on each track. Admittedly, such choices are often artistic decisions, but there are many technical tasks in the production process for which listening tests have not yet been performed to even establish whether a listener preference exists.

Listening tests must be performed to ascertain the extent to which listeners can detect undesired artifacts that commonly occur in the audio production process. Before they are ready for practical use, intelligent software tools need to be evaluated by both amateurs and professional sound engineers in order to assess their effectiveness and compare different approaches. With the exception of [20], and in contrast to separation of sources in multi-channel content, there has been little published work on subjective evaluation of the intelligent tools for mixing multi-channel audio. Where possible, prototypes should also be tested with engineers from the live sound and post-production communities in order to assess the user experience and compare performance and parameter settings with manual operation. This research would both identify preferred sound engineering approaches and allow automatic mixing criteria derived from best practices to be replaced with more rigorous criteria based on psychoacoustic studies.

4.2 An Architecture for Multichannel and Cross-Adaptive Audio Effects

One of the problems with deploying intelligent tools is that modern audio editing software does not support this architecture. Effects are single channel; single input, single output, or at most 2 in the case of stereo. Although VST3 in theory supports cross-adaptive digital audio effects, those software-based mixers that allow plug-in of VST3 effects are limited in their use. For instance, the Inserts are pre or post built in processors, thus preventing effective use of intelligent, automatic faders and equalizers. For, an 8 channel mixer with support for cross-adaptive effects was custom-built.

To address this, an automatic mixing host that supports cross-adaptive effects must be developed. An initial software prototype of such a system was built [17] for the evaluation of many of the intelligent multichannel signal processing tools described in the previous section, and integrated with a Mackie hardware control surface for demonstration. However, it was not intended to support additional cross-adaptive effects beyond those that were being developed.

Furthermore, any input or output stage in the automatic mixing host should have the capability to be analyzed, processed, stored and recalled. This would allow for portability and consistency. Hardware implementations would permit a host system to be tested in live situations. And hardware could be evolved from an automated mixer to an intelligent, automatic mixer.

5 CONCLUSIONS

In this paper, we described how mixing of multichannel audio could be made simpler and more efficient through the

use of intelligent software tools. Ideally, intelligent systems for mixing multichannel audio should be able to pass a Turing test. That is, they should be able to produce music indistinguishable from that which could be handcrafted by a professional human engineer. This would require the systems to be able to make artistic as well as technical decisions, and achieve this with almost arbitrary audio content. However, considerable progress is still needed in order for systems to even be able to 'understand' the musician's intent [21]. But, in the near term, such software tools may result in two types of systems. The first would be a set of tools for the sound engineer which automate repetitive tasks. This would allow professional audio engineers to focus on the creative aspects of their craft, and help inexperienced users create high quality mixes. The other type of system would be a 'black box' for the musician which allows decent live sound without an engineer. This would be most beneficial for the small band or small venue that don't have or can't afford a sound engineer, or for recording practice sessions where a sound engineer is not typically available.

There are major concerns with such an approach. Much of what a sound engineer does is creative, and based on artistic decisions. It is doubtful that such decisions could be effectively reproduced by a machine. But if the automation is successful, then machines may replace sound engineers. However, it is important to note that these tools are not intended to remove the creativity from audio production. Nor do they require software to reproduce artistic decisions, although this would be an interesting direction for future research. Rather, the tools rely on the fact that many of the challenges are technical engineering tasks. Some of which are perceived as creative decisions because there are a wide range of approaches without a clear understanding of listener preferences. By automating those engineering aspects of record production, it will allow the musicians to concentrate on the music and allow the audio engineers to concentrate on the more interesting, creative challenges.

6 ACKNOWLEDGMENTS

Thanks to Dr. Enrique Perez Gonzalez, Martin Morrell, Alice Clifford, Daniele Barchiesi, and others whose research was described in this paper. This work was partly funded under the EU FP7 project DigiBIC, www.digibic.eu .

7 REFERENCES

[1] P. White, "Automation for the People," *Sound on Sound,* vol. 23, October 2008.

[2] J. A. Moorer, "Audio in the New Millennium," *Journal of the Audio Engineering Society,* vol. 48, pp. 490-498, May 2000.

[3] B. McCarthy, *Sound System Design and Optimisation,* Focal Press, 2007.

[4] R. B. Dannenberg, "An Intelligent Multi-Track Audio Editor," in *ICMC*, 2007, pp. 89 - 94.

[5] P. Kleczkowski, "Method of Mixing Audio Signals and Apparatus for Mixing Audio Signals," USA Patent WO/2007/015652, 2008.

[6] N. Tsingos*, et al.*, "Perceptual Audio Rendering of Complex Virtual Environments," *ACM Transactions on Graphics (Proceedings of the SIGGRAPH-04),* vol. 23, 2004.

[7] F. Pachet and O. Delerue, "On-the-fly multi track mixing," in *109th Audio Engineering Society Convention*, 2000.

[8] 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.

[9] M. J. Terrell*, et al.*, "Automatic Noise Gate Settings for Drum Recordings Containing Bleed from Secondary Sources," *EURASIP Journal on Advances in Signal Processing,* v2010, pp. 1-9, 2010.

[10] M. Morrell and J. D. Reiss, "Dynamic Panner: An Adaptive Digital Audio Effect for Spatial Audio," in *127th AES Convention*, New York, 2009.

[11] D. Dugan, "Automatic Microphone Mixing," in *51st Convention Audio Engineering Society*, San Fransico, 1975.

[12] E. Perez Gonzalez and J. D. Reiss, "Automatic Mixing," in *Digital Audio Effects*, U. Zoelzer, Ed., 2nd ed, 2011.

[13] A. Clifford and J. D. Reiss, "Calculating Time Delays of Multiple Active Sources in Live Sound," presented at the 129th AES Convention, San Francisco, 2010

[14] C. Uhle and J. D. Reiss, "Determined Source Separation for Microphone Recordings Using IIR Filters," in *129th AES Convention*, San Francisco, 2010.

[15] M. J. Terrell and J. D. Reiss, "Automatic monitor mixing for live musical performance," *Journal of the Audio Engineering Society,* vol. 57, pp. 927-936, November 2009.

[16] D. Barchiesi and J. D. Reiss, "Reverse Engineering the Mix," *Journal of the Audio Engineering Society,* vol. 58, pp. 563-576, July 2010.

[17] E. Perez Gonzalez, "Advanced Automatic Mixing Tools for Music," PhD, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, 2010.

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

[19] 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.

[20] 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.

[21] J. S. Downie*, et al.*, "The Music Information Retrieval Evaluation eXchange" in *Advances in Music Information Retrieval* ed New York: Springer, 2010.