15

An Intelligent Systems Approach to Mixing Multitrack Audio

Joshua D. Reiss

Introduction

Although audio production tasks are challenging and technical, much of the initial work follows established rules and best practices. Yet multitrack audio content is still often manipulated ‘by hand’, using no computerized signal analysis. This is a time-consuming process, and prone to errors. Only if time and resources permit does the sound engineer refine his or her choices to produce an aesthetically pleasing mix which best captures the intended sound.

In order to address this challenge, a new form of multitrack audio signal processing has emerged. Intelligent tools have been devised that analyze the relationships between all channels in order to automate the mixing of multitrack audio content. By ‘intelligent’, we mean that these tools are expert systems that perceive, reason, learn and act intelligently. This implies that they must analyze 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 on objective and perceptual criteria. In parallel, intelligent audio production interfaces have arisen that guide the user, learn his or her preferences and present intuitive, perceptually relevant controls.

An assumption that is often, but not always, made about mixing is that it is an iterative process (Pestana, 2013). There is no fixed order in the sequence of steps applied, and an iterative, coarse-to-fine approach is applied (Figure 15.1) whereby mixing is treated as an optimization problem, with targets and criteria set for the final mix. Such a view lends itself well to an intelligent systems approach, whereby the steps can be sequenced and diverse optimization or adaptive techniques can be applied in order to achieve given objectives.

For progress towards intelligent systems in this domain, significant problems must be overcome that have not yet been tackled by the research community. First, multitrack audio editing tools demand manual intervention. Although audio editors are capable of saving a set of static scenes

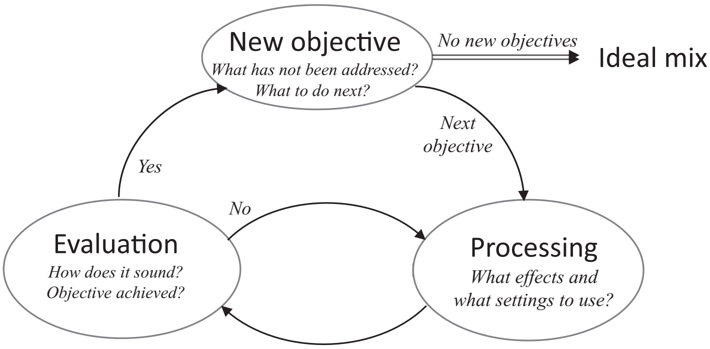

Figure 15.1 The iterative approach to mixing multitrack audio

for later use, they lack the ability to take intelligent decisions, such as adapting to different acoustic environments or different set of inputs. Second, most state-of-the-art audio signal processing techniques focus on single-channel signals. Yet multichannel or multitrack 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 analyze the content of all sources in order to improve the quality of capturing, editing and combining multitrack audio. Finally, advances in machine learning must be tailored towards problems and practical applications in the domain of audio production. This chapter presents an overview of recent advances in this area.

Enabling concepts

The idea of automating the audio production process, although relatively unexplored, is not new. In *Automation for the People* (White, 2008), 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 that mixing tools 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, Moorer (2000) 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”.

Automatic mixing research has received a lot of attention in recent years. The state of the art was described in Reiss (2011), but since then the field has grown rapidly. This section describes the key concepts in automatic mixing.

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 (Verfaille et al., 2006). 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 analyze 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 equalization, 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 Figure 15.2. Any additional processing is performed in a separate section so that the audio signal flow is unaffected. This side chain is essential for low latency, real-time signal flow. The side chain is comprised of a feature extraction section and an analysis section.

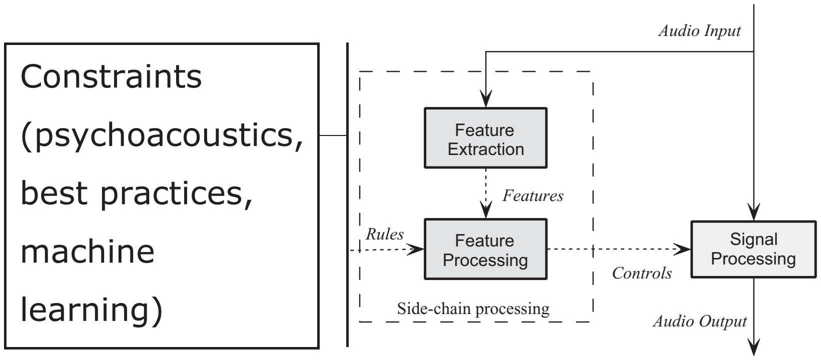

Figure 15.2 Block diagram of an intelligent audio effect. Features are extracted by analysis of the audio signal. These features are then processed based on a set of rules intended to mimic the behavior of a trained engineer. A set of controls are produced which are used to modify the audio signal

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.

Reiss (2011) described several intelligent, adaptive effects for use with single-channel audio, which automate many parameters and enable a higher level of audio editing and manipulation. 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.

Cross-Adaptive Digital Audio Effects

When editing multitrack 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 analyze 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 multitrack audio editing system, as shown in Figure 15.3, 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 multitrack 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 multitrack 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. 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.

Intelligent, Multitrack Digital Audio Effects

In Reiss (2011), and references therein, several cross-adaptive digital audio effects were described that explored the possibility of reproducing

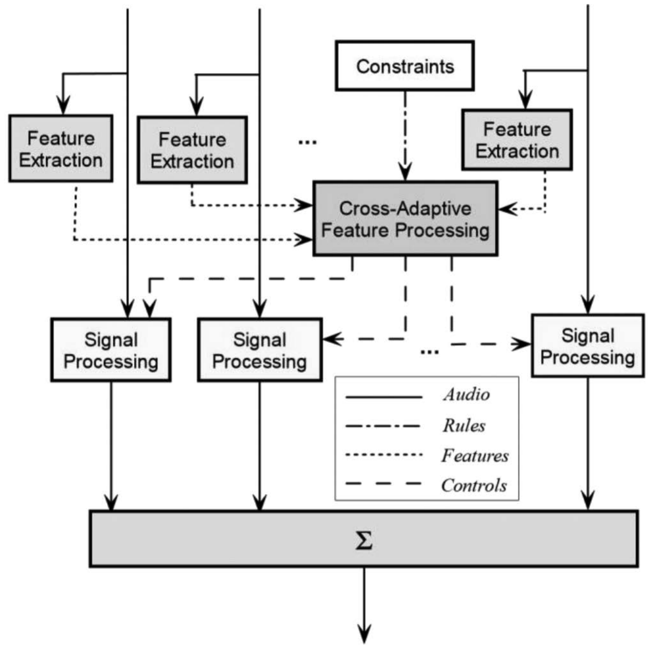

Figure 15.3 Block diagram of an intelligent, cross-adaptive mixing system. Extracted features from all channels are sent to the same feature-processing block, where controls are produced. The output channels are summed to produce a mix that depends on the relationships between all input channels

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;

$$mix_l[n] = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} c_{k,m,l}[n] * x_m[n], \tag{1}$$

where there are M input tracks and L channels in the output mix. K is the length of the control vector c and x is the multitrack input. Thus, the resultant mixed signal at time n is a sum over all input channels, of a control vectors convolved with the input signal.

Any cross-adaptive digital audio effect that employs linear filters may be described in this manner. 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 (typically, 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. If multitrack convolutional reverb is applied, then c represents direct application of a finite room impulse response. And automatic equalization employs impulse responses for the control vectors based on transfer functions representing each equalization curve applied to each channel. And though dynamic range compression is a nonlinear effect due to its level dependence, the application of feedforward compression is still as a simple gain function. So multitrack dynamic range compression would be based on a time-varying gain for each control vector.

Real-Time, Multitrack 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.

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;

$$c_m[n+1] = \begin{cases} c_m[n] & x^2_{m,RMS}[n] \le r^2_{RMS}[n] \\ \alpha c'_m[n+1] + (1-\alpha)c_m[n] & otherwise \end{cases} \qquad (2)$$

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

$$x^2_{rms}[n] = \frac{1}{M}\sum_{0}^{M-1} x^2[n-m] \qquad (3)$$

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

$$x^2_{RMS}(n+1) = x^2(n+1)/M + x^2_{RMS}(n) - x^2(n+1-M)/M\ , \qquad (4)$$

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

$$x^2_{RMS}(n+1) = \beta x^2(n+1) + (1-\beta)x^2_{RMS}(n)\ . \qquad (5)$$

α and β 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.

Incorporating Best Practices Into Constrained Control Rules

In order to develop intelligent software tools, it is essential to formalize and analyze 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 multitracks, 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 (Pestana et al., 2014b). 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 minimized, and so on.

Similarly, many aspects of sound spatialization obey standard rules. For instance, a stereo mix should be balanced and hard panning 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. Sources with similar frequency content should be placed far apart, in order to prevent spatial masking and improve the intelligibility of content. Interestingly, Wakefield et al. (2015) showed that this avoidance of spatial masking may be a far more effective way to address general masking issues in a mix than alternative approaches using equalizers, compressors and level balancing.

These best practices and common approaches translate directly into constraints that are built into intelligent software tools. For example, De Man et al. (2013a, 2013b) described autonomous systems that were built entirely on best practices found in the literature. Also, many parameters on digital audio effects can be set based on an understanding of best practices and analysis of signal content, e. g., attack and release on dynamics processors are kept short for percussive sounds.

Psychoacoustic Studies

Important questions arise concerning the psychoacoustics of mixing multitrack 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. Important work in this area has addressed issues such as level balance preference (King et al., 2010, 2012), reverberation level preference (Leonard et al., 2012, 2013), 'punch' (Fenton et al., 2015), perceived loudness and dynamic range compression (Wilson et al., 2016), as well as the design and interpretation of such listening tests.

Before they are ready for practical use, intelligent software tools need to be evaluated by both amateurs and professional sound engineers to assess their effectiveness and compare different approaches. In contrast to separation of sources in multitrack content, there has been little published work on subjective evaluation of the intelligent tools for mixing multitrack 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.

Recent developments

Table 15.1 provides an overview of intelligent mixing systems since the early ones described in Reiss (2011). These technologies are classified in terms of their overall goal, whether they are multitrack or single track, whether or not they are intended for real-time use and how their rules are found.

Many of the tools deal with masking in some form. Lopez et al. (2010), Aichinger et al. (2011) and Ma et al. (2014) all propose measures of masking in multitrack mixes, but do not contain intelligent approaches to masking reduction.

Faders

The most common form of multitrack automatic mixing system is based around simple level adjustments on each track. In almost all cases, it begins with the assumption that each track is meant to be heard at roughly equal loudness levels.

Mansbridge et al. (2012b) provided a real-time system, using ITU 1770 as the loudness model. The off-line system described in Ward et al. (2012) attempted to control faders with auditory models of loudness and partial loudness. In theory, this approach should be more aligned with perception and take into account masking, at the expense of computational efficiency. But Wichern et al. (2015) showed that the use of an auditory model offered little improvement over simple single-band, energy-based approaches. Interestingly, the evaluation in Mansbridge et al. (2012b) showed that autonomous faders could compete with manual approaches by professionals, and test subjects gave the autonomous system highly consistent ratings, regardless of the song (and its genre and instrumentation) used for testing. This suggests that the equal loudness rule is broadly applicable, whereas preference for decisions in manual mixes differs widely dependent on content.

Equalization

The rules and best practices for equalization typically fall into two categories: artifact correction, such as hum removal (Brandt et al., 2014) and the equalization of salient frequencies (Bitzer et al., 2008), or creative equalization (which may still follow rules and best practices), where equalizers are applied in order to achieve a certain overall spectrum (Pestana et al., 2013; Deruty et al., 2014).

Table 15.1 Classification of intelligent audio production tools since those described in Reiss (2011)

Single or multitrack	*Audio effect*	*Reference*	*Real-time?*	*Rules*
Single track	Equalization	Ma et al., 2013	Yes	Mix analysis
		Sabin et al., 2008, 2009a, 2009b; Pardo et al., 2012a, 2012b; Cartwright et al., 2013	No	Machine learning
	Compression	Giannoulis et al., 2013; Mason et al., 2015	Yes	Psychoacoustics; best practices
	Reverberation	Rafii et al., 2009; Chourdakis et al., 2016a	No	Machine learning
	Distortion	De Man et al., 2014b		Best practices
Multitrack	Faders and gains	Mansbridge et al., 2012b	Yes	Best practices
		Scott et al., 2011; Ward et al., 2012; Wichern et al., 2015	No	Best practices
	Equalization	Hafezi et al., 2015	Yes	Best practices
	Compression	Maddams et al., 2012; Ma et al., 2015	Yes	Psychoacoustics; best practices
	Stereo panning	Mansbridge et al., 2012a; Pestana et al., 2014a	Yes	Best practices
	Delay and polarity	Clifford et al., 2010, 2011a, 2013; Jillings et al., 2013	Yes	Acoustics
	Interference reduction	Clifford et al., 2011b; Kokkinis et al., 2011	No	Acoustics
	Exploration	Cartwright et al., 2014	Yes	Machine learning
	Knowledge engineered mix	De Man et al., 2013a, 2013b	No	Best practices

Ma et al. (2013) described an intelligent equalization tool that, in real time, equalized an incoming audio stream towards a target frequency spectrum. The target spectrum was derived from analysis of fifty years of commercially successful recordings (Pestana et al., 2013). Since the input signal to be equalized is continually changing, the desired magnitude response of the target filter is also changing (though the target output spectrum remains the same). Thus, smoothing was applied from frame to frame on the desired magnitude response and on the applied filter. Targeting was achieved using the Yule-Walker method, which can be used to design an IIR filter with a desired magnitude response.

Hafezi et al. (2015) created a multitrack intelligent equalizer that used a measure of masking and rules based on best practices from the literature to apply, in real time, different multiband equalization curves to each track. Results of objective and subjective evaluation were mixed and showed lots of room for improvement, but they indicated that masking was reduced and the resultant mixes were preferred over amateur, manual mixes.

Stereo Positioning

The premise of Mansbridge et al. (2012a) is that one of the primary goals of stereo panning is to ‘fill out’ the stereo field and reduce masking. It set target criteria of source balancing (equal numbering and symmetric positioning of sources on either side of the stereo field), spatial balancing (uniform distribution of levels) and spectral balancing (uniform distribution of content within each frequency band). It further assumes that the higher the frequency content of a source, the more it will be panned, and that no hard panning will be applied. Finally, it used a multitude of techniques to position the sources; amplitude panning, timing differences and double tracking.

Pestana et al. (2014a) took a different approach, where different frequency bands of each multitrack are assigned different spatial positions in the mix. This approach is unique among the intelligent multitrack mixing tools since it does not emulate, even approximately, what might be performed by a practitioner. That is, practitioners aim for a single position (albeit sometimes diffuse) of each source. However, it captures the spirit of many practical approaches since it greatly reduces masking and makes effective use of the entire stereo field. In fact, Matz et al. (2015) showed that dynamic spectral panning had a larger effect in the overall improvement provided by automatic mixing than any of the other tools they considered (intelligent distortion, autonomous faders and multitrack EQ).

Dynamic Range Compression

Automating dynamic range compression is much more challenging than other effects for several reasons. It is a nonlinear effect with feedback, there are complicated relationships between its parameters and its use is less understood than other effects. Nevertheless, Giannoulis et al. (2013) automated most of the parameters of a compressor such that a single parameter determines the overall amount of compression and all other parameters are optimized to the signal. This was taken one step further by Mason et al. (2015), where the amount of dynamic range compression applied is determined based on a measurement of the background noise level in the environment.

A first attempt at multitrack dynamic range compression was provided by Maddams et al. (2012). Results of evaluation were mixed, and it was difficult to identify a preference between an automatic mix, a manual mix and no compression applied at all. Furthermore, it wasn't possible to tell whether this was due to a genuine lack of preference or due to limitations in the experimental design (e.g., poor stimuli, untrained test subjects).

A more rigorous approach was taken in Ma et al. (2015). The challenge was to formalize and quantify the relevant best practices described in Pestana et al. (2014b). First, a method of adjustment test was performed to establish preferred parameter settings for a wide variety of content. Then least squares regression was used to identify the best combination of candidate features that map to parameter settings. Thus, a rule such as 'more compression is applied to percussive tracks' translates to 'the ratio setting of the compressor is a particular function of a certain measure of percussivity in the input audio track'. Perceptual evaluation then showed a clear preference for automatic dynamic range compression over amateur application and over no compression, and sometimes performed close to professionals.

Studies have also investigated the dynamic range (or loudness range) of commercial content (Deruty et al., 2014; Kirchberger et al., 2016). Though the relationship between this range and the settings of dynamic range compressors is a complicated one, this direction of research may lead the way towards automatic dynamic range compression based on matching the dynamics of popular recordings, similar to the approach taken in Ma et al. (2013) for equalization.

Delay, Polarity and Interference

Delay and interference reduction are actually well-established signal processing techniques, more generally known as time alignment and source separation, but in Clifford et al. (2010, 2011a, 2011b, 2013) and Jillings et al. (2013) they are used and customized for mixing applications. That is, they deal with optimizing parameter settings for real world scenarios, such as microphone placement around a drum kit, moving sources on stage and interference reduction under the constraint that no additional artifacts may be introduced.

Reverb

Of all the standard audio effects found on a mixing console or as built-in algorithms in a digital audio workstation, there has perhaps been the least effort on intelligent systems design for reverberation. Chourdakis et al. (2016a, 2016b) proposed an adaptive digital audio effect for artificial reverberation that allows it to learn from the user in a supervised way. They first perform feature selection and dimensionality reduction on features extracted from a training data set. Then, a user provides examples of reverberation parameters for the training data. Finally, a set of classifiers is trained, and they are compared using 10-fold cross validation to compare classification success ratios and mean squared errors. Tracks from the Open Multitrack Testbed (De Man et al., 2014a) were used in order to train and test the models.

Adaptive and Intuitive Mixing Interfaces

In this section, we provide an overview of the state of the art concerning interfaces for intelligent or adaptive mixing, with an emphasis on perceptual adaptive and intuitive controls. Various approaches for learning a listener's preferences for an equalization curve with a small number of frequency bands have been applied to research in the setting of hearing aids (Neuman et al., 1987; Durant et al., 2004) and cochlear implants (Wakefield et al., 2005), and the modified simplex procedure (Kuk et al., 1992; Stelmachowicz et al., 1994) is now an established approach for selecting hearing aid frequency responses. However, many recent innovations have emerged in the field of music production.

Dewey et al. (2013) and Mycroft et al. (2013) looked at the effect of the complexity of the interface for an equalizer, and suggested that simplified interfaces may encourage the user to focus on the aural properties of the signal, rather than the interpretation of visual information. Loviscach (2008) presented an interface for a five-band parametric equalizer, where the user simply freehand draws the desired transfer function and an evolutionary optimization strategy (chosen for real-time interaction) finds the closest match. Informal testing suggested that this interface reduced the set-up time for a parametric equalizer compared to more traditional interfaces. Building on this, Heise et al. (2010) proposed a procedure to achieve equalization and other effects using a black-box genetic optimization strategy. Users are confronted with a series of comparisons of two differently processed sound examples. Parameter settings are optimized by learning from the users' choices. Though these interfaces are novel and easy to use by the nonexpert, they make no use of semantics or descriptors.

Considerable research has aimed at the development of technologies that let musicians or sound engineers perform equalization using perceptually relevant or intuitive terms, e.g., brightness, warmth, presence. Reed (2000) presented an assistive sound equalization expert system. Inductive learning based on nearest neighbor pattern recognition was used to acquire expert skills. These are then applied to adjust the timbral qualities of sound in a context-dependent fashion. They emphasized that the system must be context dependent; that is, the equalization depends on the input signal system and hence operates as an adaptive audio effect. In Mecklenburg et al. (2006), a self-organizing map was trained to represent common equalizer settings in a two-dimensional space organized by similarity. The space was hand-labeled with descriptors that the researchers considered intuitive. However, informal subjective evaluation suggested that users would like to choose their own descriptors.

The work of Bryan Pardo and his collaborators has focused on new, intelligent and adaptive interfaces for equalization tasks. They address the challenge that complex interfaces for equalizers can prevent novices from achieving their desired modifications. Sabin et al. (2008, 2009b, 2011) described and evaluated an algorithm to rapidly learn a listener's desired equalization curve. Listeners were asked to indicate how well an equalized sound could be described by a perceptual term. After rating, weightings for each frequency band were found by correlating the gain at each frequency band with listener responses, thus providing a mapping from the descriptors to audio processing parameters. Listeners reported that the resultant sounds captured their intended meanings of descriptors, and machine ratings generated by computing the similarity of a given curve to the weighting function were highly correlated to listener responses. This allows automated construction of a simple and intuitive audio equalizer interface. In Pardo et al. (2012a), active and transfer learning techniques were applied to exploit knowledge from prior concepts taught to the system from prior users, greatly enhancing the performance of the equalization learning algorithm.

The early work on intelligent equalization based on intuitive descriptors was hampered by a limited set of descriptors with a limited set of training data to map those descriptors to equalizer settings. Cartwright et al.(2013) addressed this with SocialEQ, a web-based crowd-sourcing application aimed at learning the vocabulary of audio equalization descriptors. To date, 633 participants have participated in a total of 1,102 training sessions (one session per learned word), of which 731 sessions were deemed reliable in the sense that users were self-consistent in their answers (Pardo, 2015). This resulted in 324 distinct terms, and data on these terms is made available for download.

Building on the mappings from descriptors to equalization curves, Sabin et al. (2009a) described a simple equalizer where the entire set of curves were represented in a two-dimensional space (similar to Mecklenburg et al., 2006), thus assigning spatial locations to each descriptor. Equalization is performed by the user dragging a single dot around the interface, which simultaneously manipulates 40 bands of a graphic equalizer. This approach was extended to multitrack equalization in Cartwright et al. (2014), which provided an interface that, by varying simple graphic equalizers applied to each track in a multitrack, allowed the user to intuitively explore a diverse set of mixes.

The concepts of perceptual control, learned from crowdsourcing, intuitive interface design and mapping of a high-dimensional parameter space to a lower dimensional representation were all employed in Stasis et al. (2015). This approach scaled equalizer parameters to spectral features of the input signal, then mapped the equalizer's thirteen controls to a 2D space. The system was trained with a large set of parameter space data representing warmth and brightness, measured across a range of musical instrument samples, allowing users to perform equalization using a perceptually and semantically relevant, simple interface. A similar approach, also incorporating gestural control, was applied to dynamic range compression in Wilson et al. (2015).

Current and Future Research Directions

Open Multitrack Testbed

The availability multitrack audio is of vital importance to research in this field, but existence of such tracks alone is not sufficient. The content should be highly diverse in terms of genre, instrumentation and quality, so that sufficient data is available for most applications. Where training on large datasets is needed, such as with machine learning applications, a large number of audio samples is especially critical.

Data that can be shared without limits, because of a Creative Commons or similar license, facilitates collaboration, reproducibility and demonstration of research and even allows it to be used in commercial settings, making the testbed appealing to a larger audience.

Moreover, reliable metadata can serve as a ground truth that is necessary for applications such as instrument identification, where the algorithm's output needs to be compared to the 'actual' instrument. Providing this data makes the testbed an attractive resource for training or testing such algorithms, as it obviates the need for manual annotation of the audio, which can be particularly tedious if the number of files becomes large. Similarly, for the testbed to be highly usable, it is mandatory that the desired type of data can be easily retrieved by filtering or searches pertaining to this metadata.

Existing online resources of multitrack audio content have a relatively low number of songs, show little variation in content, contain content of which the use is restricted due to copyright, provide little to no metadata, rarely have mixed versions including the parameter settings, and/or do not come with facilities to search the content for specific criteria. However, two initiatives (Bittner et al., 2014; De Man et al. 2014a) have tried to address this problem. MedleyDB is an annotated, royalty-free dataset of multitrack recordings, initially developed to support research on melody extraction, but generally applicable to a wide range of multitrack research problems. The Open Multitrack Testbed (which also links to the MedleyDB content) was designed for broad and diverse use by researchers, educators and enthusiasts. Such initiatives are a strong indicator that research in this field will continue to grow.

Mix Evaluation

One of the chief distinguishing characteristics between the early work on intelligent mixing systems and those described herein is that very few of the early systems had any form of subjective evaluation, whereas now this is standard practice. A popular form of evaluation for such systems has become multistimulus rating, similar to that used in MUSHRA.

Mansbridge et al. (2012b) compared their proposed autonomous faders technique with a manual mix, an earlier implementation, a simple sum of sources and a semi-autonomous version. Mansbridge et al. (2012a) compared an autonomous panning technique with a monaural mix and panning configurations set manually by three different engineers. Both showed that fully autonomous mixing systems can compete with manual mixes.

Similar listening tests for the multitrack dynamic range compression system described in Maddams et al. (2012) were inconclusive, however, since the range of responses was too large for statistically significant differences between means and since no dynamic range compression was often preferred, even over the settings made by a professional sound engineer. However, a more rigorous listening test was performed in Ma et al. (2015), where it was shown that compression applied by an amateur was on a par with no compression at all, and an advanced implementation of intelligent multitrack dynamic range compression was on a par with the settings chosen by a professional.

In Wichern et al. (2015), the authors first examined human mixes from a multitrack dataset to determine instrument-dependent target loudness templates. Three automatic level balancing approaches were then compared to human mixes. Results of a listening test showed that subjects preferred the automatic mixes created from the simple energy-based model, indicating that the complex psychoacoustic model may not be necessary in an automated level setting application.

One of the most exciting and interesting developments has been perceptual evaluation of complete automatic mixing systems. In Matz et al. (2015), various implementations of an automatic mixing system are compared, where different combinations of autonomous multitrack audio effects were applied, so that one could see the relative importance of each individual tool. Although no comparison was made with manual mixes, it is clear that the application of these tools provides an improvement over the original recording, and that the combination of all tools results in a dramatic improvement.

Conclusions

In this chapter, we described how mixing of multitrack audio could be made simpler and more efficient through the use of intelligent software tools. Ideally, intelligent systems for mixing multitrack 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. 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 that 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 that allows decent live sound without an engineer. This would be most beneficial for the small band or small venue that doesn'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 is 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.

Bibliography

Aichinger, P., et al. (2011). 'Describing the Transparency of Mixdowns: The Masked-to-Unmasked Ratio.' In *130th Audio Eng. Soc Convention*.

Bittner, R., et al. (2014). 'MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research.' In *15th International Society for Music Information Retrieval Conference (ISMIR)*.

Bitzer, J., et al. (2008). 'Evaluating Perception of Salient Frequencies: Do Mixing Engineers Hear the Same Thing?' In *124th Audio Engineering Society Convention*, Amsterdam.

Brandt, M., et al. (2014). Automatic Detection of Hum in Audio Signals. *Journal of Audio Engineering Society* 62 (9): 584–595.

Cartwright, M., et al. (2013). 'Social-EQ: Crowdsourcing an Equalization Descriptor Map.' In *14th International Society for Music Information Retrieval*, Curitiba, Brazil.

Cartwright, M., et al. (2014). 'Mixploration: Rethinking the Audio Mixer Interface.' In *19th International Conference on Intelligent User Interfaces Proceedings (IUI14)*, Haifa, Israel.

Chourdakis, E.T., et al. (2016a). '*Automatic Control of a Digital Reverberation Effect using Hybrid Models*.' In *AES 60th International Conference Leuven*, Belgium.

Chourdakis, E.T., et al. (2016b). A Machine Learning Approach to Design and Evaluation of Intelligent Artificial Reverberation. *Journal of Audio Engineering Society (to appear)*.

Clifford, A., et al. (2010). 'Calculating Time Delays of Multiple Active Sources in Live Sound.' In *129th AES Convention*, San Francisco.

Clifford, A., et al. (2011a). Reducing Comb Filtering on Different Musical Instruments Using Time Delay Estimation. *Journal on the Art of Record Production* 5: 1–13.

Clifford, A., et al. (2011b). 'Microphone Interference Reduction in Live Sound.' In *14th Int. Conference on Digital Audio Effects (DAFx-11)*, Paris.

Clifford, A., et al. (2013). Using Delay Estimation to Reduce Comb Filtering of Arbitrary Musical Sources. *Journal of the Audio Engineering Society* 61 (11): 917–927.

De Man, B., et al. (2013a). 'A Knowledge-Engineered Autonomous Mixing System.' In *135th AES Convention*, New York.

De Man, B., et al. (2013b). A Semantic Approach to Autonomous Mixing. *Journal on the Art of Record Production (JARP)* 8: 1–23.

De Man, B., et al. (2014a). 'The Open Multitrack Testbed.' In *137th AES Convention*, Los Angeles.

De Man, B., et al. (2014b). 'An Intelligent Multiband Distortion Effect.' In *AES 53rd International Conference on Semantic Audio*, London, UK.

Deruty, E., et al. (2014). Human–Made Rock Mixes Feature Tight Relations between Spectrum and Loudness. *Journal of Audio Engineering Society* 62 (10): 643–653.

Dewey, C., et al. (2013). 'Novel Designs for the Parametric Peaking EQ User Interface.' In *134th AES Convention*, Rome.

Durant, E.A., et al. (2004). Efficient Perceptual Tuning of Hearing Aids with Genetic Algorithms. *IEEE Transactions on Speech and Audio Processing* 12 (2): 144–155.

Fenton, S., et al. (2015). 'Towards a Perceptual Model of "Punch" in Musical Signals.' In *139th AES Convention*, New York.

Giannoulis, D., et al. (2013). Parameter Automation in a Dynamic Range Compressor. *Journal of Audio Engineering Society* 61 (10): 716–726.

Hafezi, S., et al. (2015). Autonomous Multitrack Equalisation Based on Masking Reduction. *Journal of the Audio Engineering Society* 63 (5): 312–323.

Heise, S., et al. (2010). 'A Computer-Aided Audio Effect Setup Procedure for Untrained Users.' In *128th Audio Engineering Society Convention*, London.

Jillings, N., et al. (2013). 'Performance Optimization of GCC-PHAT for Delay and Polarity Correction under Real World Conditions.' In *134th AES Convention*, Rome.

King, R., et al. (2010). 'Variance in Level Preference of Balance Engineers: A Study of Mixing Preference and Variance Over Time.' In *129th Audio Engineering Society Convention*, San Francisco.

King, R., et al. (2012). 'Consistency of Balance Preferences in Three Musical Genres.' In *133rd AES Convention*.

Kirchberger, M., et al. (2016). Dynamic Range Across Music Genres and the Perception of Dynamic Compression in Hearing-Impaired Listeners. *Trends in Hearing* 20: 1–16.

Kokkinis, E., et al. (2011). 'Detection of 'Solo Intervals' in Multiple Microphone Multiple Source Audio Applications.' In *130th AES Convention*.

Kuk, F.K., et al. (1992). The Reliability of a Modified Simplex Procedure in Hearing Aid Frequency Response Selection. *Journal of Speech and Hearing Research* 35 (2): 418–429.

Leonard, B., et al. (2012). 'The Effect of Acoustical Environment on Reverberation Level Preference.' In *133rd AES Convention*, San Francisco.

Leonard, B., et al. (2013). 'The Effect of Playback System on Reverberation Level Preference.' In *134th Audio Engineering Society Convention*, Rome.

Lopez, S.V., et al. (2010). Quantifying Masking in Multi-Track Recordings. *Sound and Music Computing* 1–8.

Loviscach, J. (2008). 'Graphical Control of a Parametric Equalizer.' In *Audio Engineering Society Convention* 124.

Ma, Z., et al. (2013). 'Implementation of an Intelligent Equalization Tool Using Yule-Walker for Music Mixing and Mastering.' In *134th AES Convention*, Rome.

Ma, Z., et al. (2014). 'Partial Loudness in Multitrack Mixing.' In *AES 53rd International Conference on Semantic Audio*, London.

Ma, Z., et al. (2015). Intelligent Multitrack Dynamic Range Compression. *Journal of Audio Engineering Society* 63 (6): 412–426.

Maddams, J., et al. (2012). 'An Autonomous Method for Multi-Track Dynamic Range Compression.' In *Digital Audio Effects (DAFx)*, York, 1–8.

Mansbridge, S., et al. (2012a). 'An Autonomous System for Multi-track Stereo Pan Positioning.' In *133rd AES Convention*, San Francisco.

Mansbridge, S., et al. (2012b). 'Implementation and Evaluation of Autonomous Multi-Track Fader Control.' In *132nd Audio Engineering Society Convention*, Budapest, 1–8.

Mason, A., et al. (2015). 'Adaptive Audio Reproduction Using Personalised Compression.' In *AES 57th International Conference*, Hollywood, CA.

Matz, D., et al. (2015). 'New Sonorities for Early Jazz Recordings Using Sound Source Separation and Automatic Mixing Tools.' In *ISMIR*, Malage.

Mecklenburg, S., et al. (2006). 'subjEQt: Controlling an Equalizer through Subjective Terms.' In *Computer-Human Interaction EA '06 Extended Abstracts on Human Factors in Computing, Montreal*.

Moorer, J.A. (2000). Audio in the New Millennium. *Journal of the Audio Engineering Society* 48 (5): 490–498.

Mycroft, J., et al. (2013). 'The Influence of Graphical User Interface Design on Critical Listening Skills.' In *Sound and Music Computing (SMC)*, Stockholm.

Neuman, A.C., et al. (1987). An Evaluation of Three Adaptive Hearing Aid Selection Strategies. *The Journal of Acoustical Society of America* 82 (6): 1967–1976.

Pardo, B. (2015). *Data on SocialEQ*. J.D. Reiss.

Pardo, B., et al. (2012a). 'Building a Personalized Audio Equalizer Interface with Transfer Learning and Active Learning.' In *2nd International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM)*, Nara, Japan.

Pardo, B., et al. (2012b). 'Towards Speeding Audio EQ Interface Building with Transfer Learning.' In *New Interfaces for Musical Expression*, Ann Arbor, MI.

Pestana, P. (2013). *Automatic Mixing Systems Using Adaptive Audio Effects.* PhD, Universidade Catolica Portuguesa.

Pestana, P.D., et al. (2013). 'Spectral Characteristics of Popular Commercial Recordings 1950–2010.' In *135th AES Convention*, New York.

Pestana, P.D., et al. (2014a). 'A Cross-Adaptive Dynamic Spectral Panning Technique.' In *17th Int. Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany.

Pestana, P.D., et al. (2014b). 'Intelligent Audio Production Strategies Informed by Best Practices.' In *AES 53rd International Conference on Semantic Audio*, London, UK.

Rafii, Z., et al. (2009). 'Learning to Control a Reverberator using Subjective Perceptual Descriptors.' In *10th Int. Conf. Music Inf. Retrieval (ISMIR)*, Kobe, Japan.

Reed, D. (2000). 'A Perceptual Assistant to Do Sound Equalization.' In *5th International Conference on Intelligent User Interfaces*, New Orleans.

Reiss, J.D. (2011). 'Intelligent Systems for Mixing Multichannel Audio.' In *17th International Conference on Digital Signal Processing (DSP2011)*, Corfu, Greece, 1–6.

Sabin, A., et al. (2008). 'Rapid Learning of Subjective Preference in Equalization.' In *125th Audio Engineering Society Convention*, San Francisco.

Sabin, A., et al. (2009a). '2DEQ: An Intuitive Audio Equalizer.' In *ACM Creativity and Cognition*, Berkeley, CA.

Sabin, A., et al. (2009b). 'A Method for Rapid Personalization of Audio Equalization Parameters.' In *ACM Multimedia*, Beijing, China.

Sabin, A.T., et al. (2011). Weighted-Function-Based Rapid Mapping of Descriptors to Audio Processing Parameters. *Journal of Audio Engineering Society* 59 (6): 419–430.

Scott, J., et al. (2011). 'Automatic Multi-Track Mixing Using Linear Dynamical Systems.' In *8th Sound and Music Computing Conference (SMC)*, Padova.

Stasis, S., et al. (2015). 'A Model for Adaptive Reduced-Dimensionality Equalisation (Best Paper, 2nd Prize).' In *18th Int. Conference on Digital Audio Effects (DAFx-15)*, Trondheim, Norway.

Stelmachowicz, P.G., et al. (1994). Preferred Hearing-Aid Frequency Responses in Simulated Listening Environments. *Journal of Speech and Hearing Research* 37 (3): 712–719.

Verfaille, V., et al. (2006). Adaptive Digital Audio Effects (A-DAFx): A New Class of Sound Transformations. *IEEE Transactions on Audio, Speech and Language Processing* 14 (5): 1817–1831.

Wakefield, G.H., et al. (2005). Genetic Algorithms for Adaptive Psychophysical Procedures: Recipient-Directed Design of Speech-Processor MAPs. *Ear Hear* 26 (4): 57S–72S.

Wakefield, J., et al. (2015). 'An Investigation into the Efficacy of Methods Commonly Employed by Mix Engineers to Reduce Frequency Masking in the Mixing of Multitrack Musical Recordings'. In *138th AES Convention*.

Ward, D., et al. (2012). 'Multi-Track Mixing Using a Model of Loudness and Partial Loudness.' In *Audio Engineering Society (AES) 133rd Convention*, San Francisco.

White, P. (2008). Automation for the People. *Sound on Sound* 23 (12).

Wichern, G., et al. (2015). 'Comparison of Loudness Features for Automatic Level Adjustment in Mixing.' In 139th AES Convention, New York.

Wilson, T., et al. (2015). 'A Semantically Motivated Gestural Interface for the Control of a Dynamic Range Compressor.' In *138th AES Convention*.

Wilson, A., et al. (2016). Perception of Audio Quality in Productions of Popular Music. *Journal of Audio Engineering Society* 64 (1/2): 23–34.