INSTRUMENT IDENTIFICATION INFORMED MULTI-TRACK MIXING

Jeffrey Scott and Youngmoo E. Kim
Music and Entertainment Technology Laboratory (MET-lab)
Electrical and Computer Engineering, Drexel University
{jjscott, ykim}@drexel.edu

ABSTRACT

Although digital music production technology has become more accessible over the years, the tools are complex and often difficult to navigate, resulting in a large learning curve for new users. This paper approaches the task of automated multi-track mixing from the perspective of applying common practices based on the instrument types present in a mixture. We apply basic principles to each track automatically, varying the parameters of gain, stereo panning, and coarse equalization. Assuming all instruments are known, a small listening evaluation is completed on the mixed tracks to validate the assumptions of the mixing model. This work represents an exploratory analysis into the efficacy of a hierarchical approach to multi-track mixing using instrument class as a guide to processing techniques.

1. INTRODUCTION

The pervasive use of digital tools for creating, recording, producing and editing audio has led to a desire for increased automation and efficiency of these tools. Although there is a wide variety of digital audio workstations (DAW) and plug-in suites available, the level of expertise required to operate them proficiently necessarily inhibits many newcomers from obtaining reasonable results even with a significant amount of effort. This has led to an exploration in the audio signal processing community for methods of automatically analyzing audio and improving the perceived quality. Several significant difficulties arise when attempting this task. The qualitative difference between the preference of individuals, the wide range of timbre, dynamics and instrumentation and the multitude of production techniques available present ample hurdles to overcome.

This paper attempts to exploit some of the commonalities between processing chains for a specific instrument class, namely the drum kit. Figure 1 shows the general framework for the task. Given the identities of individual drum tracks (kick, snare, tom and overhead), we apply some basic guidelines to make the kit sound more balanced using spatial and spectral modifications.

In order to apply the processing techniques, the labels of each drum track must be known. To this end, we provide a simple classification experiment to evaluate the difficulty of classifying these tracks in a real world situation. A listening test is conducted to determine how well the model mixes the individual drum tracks. For the listening test, we use the ground truth instrument labels to evaluate how well the instrument based model mixes the tracks.

The remainder of the paper is organized as follows. Prior work in automatic mixing and instrument identification is presented in Section 2. Details about the dataset are outlined in Section 3 and the processing employed to form a drum mix is in Section 4. Sections 5 and 6 detail the classification experiment and listening evaluation, respectively.

2. BACKGROUND

Interest in automating multi-track production tasks has increased in recent years, focusing on both real-time and offline models based on perceptual information, acoustic features and best practices [1, 2].

Significant work on cross-adaptive methods for multi-track mixing for specific parameters has been explored in [3–7]. The implementation of these models relies on computing time domain and spectral features on the tracks present in a mixture and devising models to control mixing parameters based on the temporal and inter-track relations between the features. Psychoacoustic models of loudness and frequency masking are leveraged to inform the relationships on a perceptual level and more closely approximate the signal characteristics that are processed by humans.

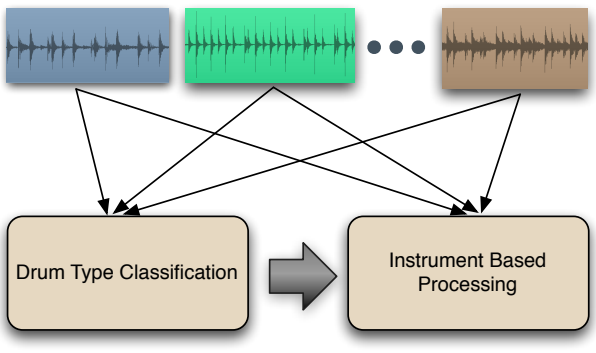

Figure 1: Identification informed mixing.

Related work focuses on estimating production techniques from user examples as well as higher level constructs of musical timbre. Variation in the vocabulary used to describe music as well as individual artist and listener preference makes universal rules difficult to derive. Observing and recording user interaction with mixing tools creates data that can be used to model individual preferences [8–10].

There have also been developments in learning mixing parameters from data. The authors in [11] use regression to estimate fader gains, equalization and compression from mixed audio signals. In a previous paper, we attempted to learn dynamic mixing models directly from data [12].

Instrument identification is a problem that has been popular in Music-IR for many years. State of the art performance on datasets of individual samples of both pitched and un-pitched instruments has reached classification accuracies in the upper 90th percentile [13–15]. However, most of these systems evaluate libraries of single, isolated instrument tones. There has been some work that shows good performance on longer excerpts of pitched instruments [16], but there are not comparable results to individual sample classification.

3. DATASET

The dataset used throughout this paper consists of 135 songs across a variety of genres. The genres include Acoustic, Alternative, Country, Dance, Electronic, Hip-Hop, Indie, Jazz, Rock and Metal. The songs were obtained from three primary sources: Weathervane Music [1], Sound on Sound [2] and a multi-track dataset used for song structure segmentation [17]. Each track is converted to a monaural source at 44.1kHz sampling rate and labeled with the instrument present in the track.

The tracks in every song are labeled with the instrument present by three individuals and the majority label for each track was retained as ground truth. The labelers are students in the music industry program at Drexel University. The filenames for each audio track are used when possible and normalized to a standard label for a single instrument class. Instrument classes are differentiated on a fine level (clean/distorted electric guitar) and may be combined into superclasses (electric guitar) if desired. The electric guitar is a specific example where fine level labels are desired since the distorted and clean versions are treated very differently by engineers and have much different roles in the mix. The dataset is publicly available online [3].

4. INSTRUMENT BASED PROCESSING

In this approach we attempt to codify some common practices and apply them to multi-track drum audio. Several professional and student mixing engineers were interviewed about the process of mixing audio and it was unsurprising to find that all of them specified that their approach is dependent upon the source material (i.e. genre, instrumentation). It is quite difficult to define a set of hard and fast rules for mixing audio yet there do exist some commonalities that many agree upon. We apply some basic techniques to improve the balance and quality of the drums via stereo panning, filtering and level adjustment. The motivation for the processing techniques employed in the following subsections are derived from the engineer interviews as well as authoritative sources on mixing [18,19].

There are several concerns when combining the signals from multiple drum microphones to produce a mixture. Problems with phase coherence between the different microphones can often occur and result in a comb filtering effect applied to the instruments [18]. This is the case with bleed (leakage) between microphones on different instruments as well as multiple microphones on a single instrument (as in the top/bottom heads of a snare drum). In properly recorded material this effect is usually anticipated for and dealt with during signal capture and therefore not considered in this paper.

We consider three proccessing areas: level balancing, stereo panning and equalization. Two basic approaches for level adjustment are serial (faders down) and parallel (faders up) [18,19]. The serial approach invovles adding in layers one at a time and the parallel approach starts with all layers active and adjusts levels accordingly. We opt for the parallel approach where the level of each instrument track is evaluated individually against the rest of the mix. There are also two main approaches to using the ambient (overhead/room) mics. One primarily uses the overheads as the main drum signal and uses the individual instrument mics as reinforcement when needed. The alternate approach is to use the close microphones as the primary signal source and use the overhead microphones to increase the amount of cymbals and add 'air' to the mix. We opt to use the latter approach in this work.

For panning, one may start with a stereo spread of the overhead mics and pan the close microphones according to their position in that signal. Another common approach is to pan the kick and snare dead center since they are the driving force of the rhythm section. This is the option we choose in our model.

The equalization applied is minimal and was obtained from the interviews of engineers. The interviewees expressed reservation about making gereralizations without hearing the source material and knowing what other instruments are in the mixture, yet these are the same issues they expressed with nearly all aspects of mixing, namely that each session is different and must be approached individually. Nevertheless, a filtering scheme was developed to boost frequency ranges that often need boosting and cut frequency ranges that often need attenuating. Ideally, this would be done adaptively through comparing bandwise energy ratios and making adjustments accordingly.

Before processing, each track is analyzed to determine where the instrument is playing on each track. We only want to compare signal characteristics where there is an ac-

[1] http://weathervanemusic.org/

[2] http://www.soundonsound.com/

[3] http://music.ece.drexel.edu/research/AutoMix

tive instrument in a track, not where there is just the noise floor. Figure 2 depicts the computation of the active regions in each track. The first four steps, full-wave rectification, low pass filtering, downsampling and smoothing with a moving average filter produce the temporal envelope of the signal and the threshold determines active regions. After thresholding, any segments less than 150ms long are discarded.

4.1 Stereo Panning

In [5, 20] a dynamic cross-adaptive model is used to actively pan tracks as they come in and out of the instrument mixture based on several constraints related to spectral and spatial balance and masking. Here we attempt to leverage common practices in drum kit panning and apply them to the individual tracks of a drum kit. This results in a static value being applied to the entire track for the duration of the song regardless of the presence or absence of instrument playing at any given time. Panning a drum kit is one aspect of mixing that is fairly consistent between engineers. Qualitatively, the stereo balance of the drum mix is as follows:

1. Kick drum panned center
2. Snare drum panned center
3. Toms panned from left to right
4. Overhead microphones panned left and right

Panning is accomplished by applying the sine-cosine panning law

$$L_{pan} = cos(45^\circ - \theta) \quad (1)$$

$$R_{pan} = sin(45^\circ - \theta). \quad (2)$$

Here $\theta \in [-45^\circ, 45^\circ]$ and represents the angle offset from the center of the stereo field with -45° being panned fully to the left and $+45^\circ$ panned fully right. This method of panning maintains the perceived loudness of the signal as it is varied from left to right. Table 1 shows the parameter values used to pan the tracks.

The kick and snare drums are panned in the center of the stereo field. The toms are spaced linearly from left to right with 25° being the maximum offset from the center position. The overhead tracks are panned alternating left and right at the specified value in Table 1.

4.2 Relative Levels

After panning, the loudness of each track is computed and compared against the loudness of the rest of the tracks to determine any boost or attenuation that is desired for each track. The loudness of each track is calculated by filtering the signal using the inverse of the ISO 226 normal equal-loudness-level contours (at 75 phons) and then computing the RMS energy over a 23ms window [21]. The level of 75 phons was chosen based on preferred listening levels shown in [22]. The loudness of the target track ($\mathbf{x}_{loud}$) is compared to the loudness of the sum of the remaining tracks ($\mathbf{y}_{loud}$) and a loudness ratio is computed,

$$r_{loud} = \frac{1}{T}\sum_{\tau}^{T} \frac{x_{loud}^{(\tau)}}{y_{loud}^{(\tau)}}, \quad (3)$$

where x and y are in dB and T is the total number of short time frames in the current song being analyzed. The loudness ratio is then used to attenuate or boost the level of the track in question. The gain of the track is determined using the following equation

$$g = 10^{(-\frac{1}{\lambda} log(r_{loud}))}. \quad (4)$$

Equation 4 offers control over the amount of level correction that is applied to each instrument through the parameter λ. As λ increases, the amount of level correction is reduced as shown in Figure 3.

Loudness is computed on each channel (L/R) after panning and the average of the loudness ratios is used to determine the gain of the instrument. There are three parameters $\{\alpha, \beta, \lambda\}$ for each instrument type that determine how the loudness ratio affects the gain, g, applied to the track. The α and β parameters define thresholds for the loudness ratio necessary to apply loudness correction. For example, if we require $r_{loud} < \alpha$ or $r_{loud} > \beta$ where $\alpha = 0.8$ and $\beta = 1.2$ before applying gain g, then the track will have no level correction if $r_{loud} \in [0.8, 1.2]$ and will have loudness correction specified in Equation 4 otherwise. The parameters in Table 1 are specified to err on the side of more kick and snare drum than overhead and tom microphones since the kick and snare instruments are generally more prominent in rock music.

Instrument Class	**Panning Value (θ)**	**Gain Values $\{\alpha, \beta, \lambda\}$**
Kick Drum	0 (center)	{0.9, 1.2, 2}
Snare Drum	0 (center)	{0.9, 1.2, 2}
Toms	Spaced {-25, 25}	{0.8, 1.3, 4}
Overhead/Room	{-35, 35}	{0.8, 1.3, 4}

Table 1: Mixing parameter values for individual drum tracks.

4.3 Equalization

The desired frequency content for a specific instrument is very genre dependent. For example in an electronic track the kick drum generally contains more low frequency content and may be prominent even into the sub-bass range. In heavy metal, the sound of the beater striking the kick drum is often desirable and the signal may need to be boosted in

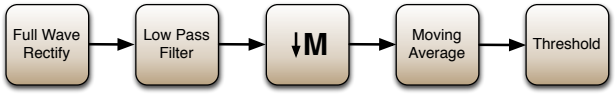

Figure 2: Processing chain to calculate the active areas of an instrument track.

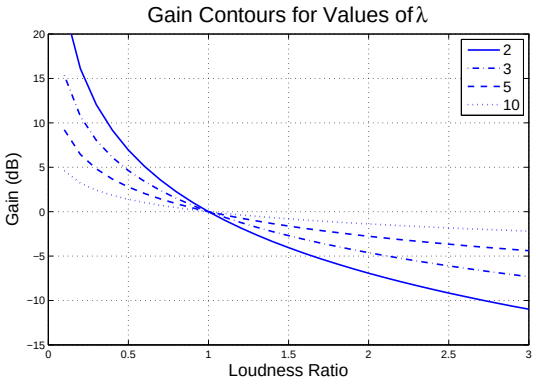

Figure 3: Contours of gain attenuation for various γ.

the high-mid frequency range. For these reasons we chose to apply only subtle equalization based on some common operations. The kick drum has a 2dB boost from 1kHz-6kHz, a 2dB cut from 400Hz-900Hz and a 2dB boost of 100Hz with a quality factor of 4.5. The snare drum has a 3dB high shelving boost starting at 10kHz. These modifications are designed to give the kick drum slightly more punch and the snare drum more brilliance.

5. DRUM TYPE CLASSIFICATION

For an unknown set of tracks, the drums would need to be identified to apply the common practices outlined above. Here we explore a preliminary experiment to classify a track in terms of the drum content it contains. The approach is fairly standard for supervised learning and is meant to serve as a benchmark of the difficulty of this particular dataset.

A support vector machine (SVM) classifier with radial basis function (RBF) kernel is trained and evaluated via 5-fold cross validation using LIBSVM [23] . This is a four class problem ($C \in \{1, 2, 3, 4\}$) with the four classes being kick drum, snare drum, tom-tom and overhead. The features used in the experiment are listed in Table 2 and include mel-frequency cepstral coefficients (MFCC) (20 dimensions), spectral features and time domain features as well as information about the amount of time active audio is present in the track. The first and second derivatives of each feature (non-singleton) is also included in the dataset. This results in 138 total feature dimensions which is then reduced through principle components analysis (PCA). The classifier achieved an average accuracy across all folds of 0.504. For a four class problem, this

Features	Features (cont.)
MFCC	RMS
Centroid	Bandwidth
Flux	Zero-Crossing Rate
Number of Segments	Inter Onset Interval
Segment Length	

Table 2: Features used in drum type classification.

Production Familiarity	Participants
None	4
Novice	4
Intermediate	6
Expert	1

Table 3: Listening test participant familiarity with audio mixing and production.

result is not particularly promising, but the model and features used are not as advanced as those in [13–16]. Although the data is in multi-track format, there are still several instruments present via the bleed of the microphones. For the tom-tom drums, the majority of the track resembles an overhead microphone signal of low amplitude until the drum is (with relative infrequency) struck. This type of real-world situation increases the difficulty of performing classification.

6. LISTENING EVALUATION

To evaluate the ability of the model to appropriately mix the drum tracks together, a listening test is performed where participants noted their preference for the individual monaural tracks summed versus the mix generated with the model. The ground truth instrument labels are used for generating the mixes using the model. Ten songs were selected at random from the dataset and a 15 second clip for each song was selected so that as many of the individual drum tracks were active as possible. Most songs in the dataset do not have drum stems associated with them, only the raw unmixed multi-track session and the final professional mix. The majority of songs that do have mixed drum stems are from the same studio and use very similar processing chains. Therefore to avoid over-representing that subset we only included the summed mix and the automatic mix across a larger number of sources.

The clip pairs were presented with the summed version and the automatically mixed version appearing in random order. Each participant was presented clip pairs one at a time and asked which clip sounds more balanced. They

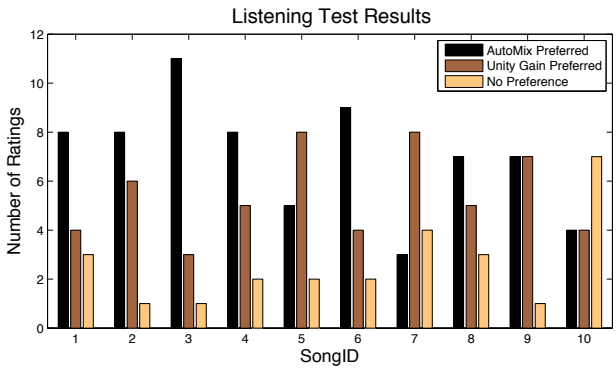

Figure 4: Listening test results showing the number of ratings for each clip pair.

could choose Clip A, Clip B or No Preference. The participants were asked to provide their level of experience with audio mixing and production, the distribution is shown in Table 3. Subjects are graduate and undergraduate students at Drexel in the music industry and engineering programs. Most subjects are male, with only two participants being female. There were 15 total participants in the study with about half having little experience working with audio production and the other half having significant experience.

Figure 4 shows the results of the listening test. For six of the ten songs, the model is preferred over the summed mix and listeners prefer two of the ten monaural summed mixes. Songs 7, 8 and 10 contain some drum loops from a library and do not adhere to the 'standard' recording technique of having kick, snare, tom and overhead microphones. The dataset represents a variety of material from various sources and varying quality. Some material is recorded professionally and sounds reasonably balanced through just summing the tracks.

7. DISCUSSION AND FUTURE WORK

We present an approach to mixing multi-track audio in an automated fashion by incorporating common techniques based on prior knowledge of instrumentation. The method obtains fair performance on a certain class of song in the dataset but is not able to gracefully handle inconsistencies in recording quality present the dataset. One caveat of working with multi-track audio is the lack of standardization for recording sessions. This makes obtaining well labeled consistent datasets to train models a difficult task in itself. The work here demonstrates the possible potential of a hierarchical system that combines both best practices and common techniques of mixing engineers with more sophisticated models of instrument identification, however there is significant room for improvement.

For the classification task, there exist more advanced methods in the literature, yet most apply to individual instrument samples and not full recorded tracks. Including more information about the temporal evolution of a signal as well as taking advantage of the audio in multiple drum tracks while classifying each track could improve results significantly.

Genre information plays a significant role in the desired drum sound for a given song. A jazz kit requires much different treatment than a dance or house drum beat, however genre recognition is not a solved problem and the definitions of genres are constantly evolving. This is an aspect of automatic mixing where it would make sense to expose a parameter to the user and offer 'presets' similar to most audio plugins.

More adaptive methods can be used on the track level processing that computes the active segments and loudness as in [7]. Perhaps the most important aspect is further user evaluation and iteration based on listening test results. The ultimate goal of automated mixing systems is to make the mix sound better to the user. Mixing audio often demands an iterative coarse-to-fine approach where the engineer is constantly making changes and then evaluating those decisions in the context of the mix [18, 19].

This is an introductory work that explores the potential of a hierarchical approach to multi-track mixing using instrument class as a guide to processing techniques. While the classification and listening evaluation results have room for improvement, a system basing mixing decisions on the instruments in the mixture warrants further investigation.

8. REFERENCES

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

[2] E. Perez-Gonzalez and J. D. Reiss, "Automatic mixing," in *DAFX: Digital Audio Effects*, 2nd ed., U. Zölzer, Ed. John Wiley & Sons, Ltd, 2011, pp. 523–549.

[3] E. Perez-Gonzalez and J. D. Reiss, "Automatic gain and fader control for live mixing," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009.

[4] ——, "Automatic equalization of multichannel audio using cross-adaptive methods," in *127th AES Convention*, 2009.

[5] S. Mansbridge, S. Finn, and J. D. Reiss, "An autonomous system for multitrack stereo pan positioning," in *133rd AES Convention*, 2012.

[6] ——, "Implementation and evaluation of autonomous multi-track fader control," in *132nd AES Convention*, 2012.

[7] D. Ward, J. D. Reiss, and C. Athwal, "Multitrack mixing using a model of loudness and partial loudness," in *133rd AES Convention*, Oct. 2012.

[8] A. T. Sabin and B. Pardo, "A method for rapid personalization of audio equalization parameters," *Proceedings of ACM Multimedia*, pp. 769–772, 2009.

[9] Z. Rafii and B. Pardo, "Learning to control a reverberator using subjective perceptual descriptors," in *Proceedings of the 10th International Society for Music Information Retrieval Conference*, Kobe, Japan, October 26-30 2009, pp. 285–290.

[10] B. Pardo, D. Little, and D. Gergle, "Building a personalized audio equalizer interface with transfer learning and active learning," in *Proceedings of the 2nd International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*. New York, USA: ACM, 2012, pp. 13–18.

[11] D. Barchiesi and J. Reiss, "Reverse engineering of a mix," *Journal of the Audio Engineering Society*, vol. 58, no. 7, pp. 563–576, 2010.

[12] J. Scott, M. Prockup, E. M. Schmidt, and Y. E. Kim, "Automatic multi-track mixing using linear dynamical systems," in *Proceedings of the 8th Sound and Music Computing Conference*, Padova, Italy, 2011.

[13] S. Scholler and H. Purwins, "Sparse approximations for drum sound classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 933–940, 2011.

[14] A. Tindale, A. Kapur, G. Tzanetakis, and I. Fujinaga, "Retrieval of percussion gestures using timbre classification techniques," in *Proceedings of the 5th International Society for Music Information Retrieval Conference*, 2004.

[15] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.

[16] S. Essid, G. Richard, and B. David, "Musical instrument recognition by pairwise classification strategies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1401–1412, 2006.

[17] S. Hargreaves, A. Klapuri, and M. Sandler, "Structural Segmentation of Multitrack Audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2637–2647, 2012.

[18] M. Senior, *Mixing Secrets for the Small Studio*, 1st ed. Focal Press, 2011.

[19] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools*, 1st ed. Elsevier Ltd., 2008.

[20] E. Perez-Gonzalez and J. D. Reiss, "A real-time semi-autonomous audio panning system for music mixing," *EURASIP Journal on Advances in Signal Processing*, 2010.

[21] International Standards Organization, *ISO226: Normal equal-loudness-level contours*, 2003.

[22] W. E. Hodgetts, J. M. Rieger, and R. A. Szarko, "The effects of listening environment and earphone style on preferred listening levels of normal hearing adults using an mp3 player," *Ear and Hearing*, vol. 28, no. 3, pp. 290–297, 2007.

[23] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.