Audio Engineering Society

Convention e-Brief 526

Presented at the 147th Convention
2019 October 16 – 19, New York

Exploring preference for multitrack mixes using statistical analysis of MIR and textual features

Joseph Colonel[1] and Joshua Reiss[1]

[1]*Queen Mary University of London*

Correspondence should be addressed to Joseph Colonel (j.t.colonel@qmul.ac.uk)

ABSTRACT

We investigate listener preference in multitrack music production using the Mix Evaluation Dataset, comprised of 184 mixes across 19 songs. Features are extracted from verses and choruses of stereo mixdowns. Each observation is associated with an average listener preference rating and standard deviation of preference ratings. Principal component analysis is performed to analyze how mixes vary within the feature space. We demonstrate that virtually no correlation is found between the embedded features and either average preference or standard deviation of preference. We instead propose using principal component projections as a semantic embedding space by associating each observation with listener comments from the Mix Evaluation Dataset. Initial results disagree with simple descriptions such as "width" or "loudness" for principal component axes.

Introduction

Though often overlooked by a lay audience, the mix engineer is a crucial player in executing a musician's vision [1]. A mixing engineer is expected to maintain expert knowledge of how to apply digital processing to audio, utilize equipment in a studio, treat sonic elements in a song so that each stand out, and countless other things. However, the advent of cheap computing and the digital audio workstation (DAW) has given amateur musicians and mixers access to many of the tools and workflows professionals use, with little to no guidance [2]. There is a clear and present need to provide tools to these mixers that can help them mix more professionally.

The first step in helping guide amateur mixers requires a model of professional mixing behaviour. The question remains, though, of how to model this behaviour. Over the past decade much research has been published to address portions of this quandry, such as interrogations of mixing "best practices and common sense." For example, [3] demonstrates that suggestions in mixing literature often conflict with the practice of professional mixers. Another algorithm, presented in [4], uses least-squares optimization techniques that can estimate processing such as panning position and gain envelopes to reverse-engineer how a track was mixed.

Furthermore, much work has been published on algorithms for autonomous and assistive mixing [5]. These include black box algorithms for tonal balance enhancement [6], algorithms to properly group and panning percussive stems of a song [7], and plugins that aid users by providing a map of semantic descriptors to effect settings [8]. These approaches do not attempt to model the whole mixing process, however, and thus do not lend themselves to complete characterization of mixing behaviour.

One approach attempts to answer this behaviour mod-

elling question by using feature extraction from mixdowns, and subsequently characterizing the dimenions of a Principal Component Analysis (PCA) [9] (based on previous work [10]). This analysis uses a combined 1501 mixes of 10 songs from the Cambridge Multitracks, performing feature extraction on 30 seconds of each mix's chorus. The scraped data was somewhat inconsistent, however. Mixers uploaded their work with varying audio quality, and the choruses were not guaranteed to be the same length or use the same stems across mixes. Furthermore, the semantic descriptors provided by the authors of the principal component axes were based on a cursory reading of the PCA axis loadings, and not in a more grounded methodology. Finally, [9] established no framework for evaluating listener preference.

The work presented here couples the feature extraction and PCA of [9] with a standardized mixing dataset that contains text data evaluating each mix in the dataset. Furthermore, listener preference ratings were collected for each mix. Results disagree with two of [9]'s assertions: that preference for a mix can be correlated with PCA embeddings, and that principal component axes describe aspects of a mix such as "balance" and "loudness."

Methods

Dataset and Feature Extraction

The Mix Evaluation Dataset (MED) is comprised of 192 mixes, and approximately 5000 evaluations of these mixes [11]. A total of 19 songs were mixed by various university groups globally. Each group was presented with the same set of stems for each song, and were instructed not to alter the stems beyond level, panning, and effects processing. This ensured that each mix of a given song contained the same content, and only varied in mixing style. Participants were then asked to rate their preference for, and comment on, each mix.

In contrasat to [9] and [10], where features were extracted only from a chorus of each mix, this analysis extracts 33 features from each verse and chorus in the MED. This is done to capture the differences that may occur between a verse and chorus, such as loudness or additional instrumentation, and adds an extra richness to the dataset. Thus a total of 384 observations were generated for principal component analysis. The features can be divided into roughly four groups, measuring the spectral characterstics, sub-band flux, loudness, and probability mass function (PMF) of the signal. The PMF is calculated by taking a histogram of the values a signal takes and normalizing. Statistical measures of this distribution are taken and can be used to characterize the distortion profile of a mix [12]. A full accounting of each feature can be found in Table 1.

Each mix was rendered as a 192kbps mp3 with 48kHz sampling rate. Each verse and chorus was downsampled to 44.1kHz for processing. All audio loading and processing was handled by the Essentia library in Python [15]. Principal component analysis was performed on these 384 observations using the sklearn library in Python [17].

Table 1: *List of extracted features.*

Feature Name	Reference
Crest Factor 100ms & 1s	-
Sub-Band Flux 0-9	[13]
EBU R128 Loudness Measures	[14]
PMF Kurtosis	[12]
PMF Skew	[12]
PMF Centroid	[12]
PMF Spread	[12]
Spectral Kurtosis	[15]
Spectral Skew	[15]
Spectral Centroid	[15]
Spectral Spread	[15]
Spectral Entropy	[15]
Spectral Rolloff 85% & 95%	[15]
Stereo Panning Spectrum 0-3	[15]
L/R Balance	[16]
Side-Mid Ratio	[16]

Principal Component Analysis

Figures 1 and 2 show how the features load the first three axes of the principal component analysis. These first three principal components explain a total of 62.7% of the data's variance (24.3%, 23.0%, and 15.4% resepctively). Their singular values are 1.04, 1.01, and 0.83 respectively. Three components (rather than the four proposed in [9]) were chosen for visualization purposes.

Of note in this analysis is how the feature loadings differ from [9], specifically how the PMF related features dominate the first principal component. This suggests

that the PMF features account for much of the variation in the MED. The authors suspect that these features dominate because the MED includes genres such as jazz and folk alongside rock music, whereas the dataset used in [9] drew only from rock, punk, and metal music. Distortion and fuzz effects are expected to be found in punk and metal mixes, but not necessarily expected in a jazz arrangement with piano and string accompaniment. This would mean that values such as PMF skew or kurtosis would be more varied across the MED than in [9]'s dataset, as distortion would be less uniformly applied across the songs of the MED. These results support the notion that PMF related features are key to characterizing mixing behaviour, especially when considering what techniques are applied across genres.

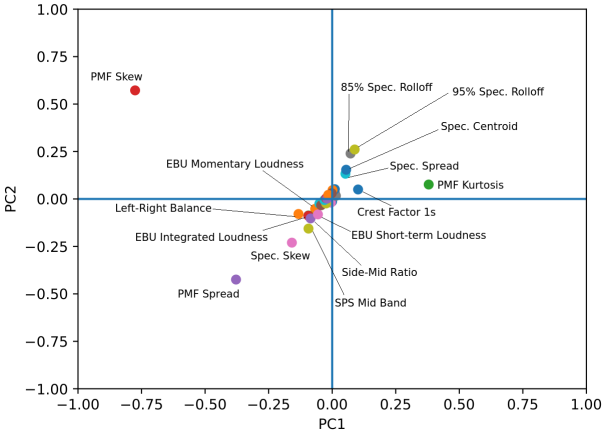

Fig. 1: *Feature loadings on the first and second principal component axes. PMF features strongly load the first axis, and spectral features strongly load the second axis.*

Preference Rating and Listener Agreement

Listeners who took part in the creation of the MED commented on each mix as well as assigning each mix a preference rating from 0 to 1, with 0 indicating a mix they disliked, and 1 indicating a mix they liked very much. Thus for each observation mentioned in Section 2.1, an average preference rating and standard deviation of preference rating is assigned. The average rating describes some notion of how "good" a mix

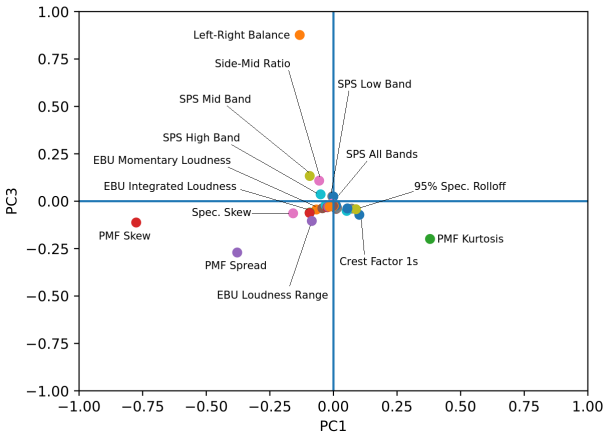

Fig. 2: *Feature loadings on the first and third principal component axes. Stereo features strongly load the third axis.*

is, and the standard deviation of preference rating describes whether listeners agreed or disagreed in their assessment (distinct from whether a mix is "good" or "bad").

Results

Linear correlations were separately calculated between the values observations took in each prinicpal component and both average listener preference and standard deviation of preference. Furthermore, correlations were calculated for the square value in each principal component and both average listener preference and standard deviation of preference. Finally, a linear correlation was calculated between the radius of each observation from the origin of the PCA embedding space and both average listener preference and standard deviation of preference. Results are shown in in Tables 2 and 3.

Discussion

The results of the embedding and correlations show that no clear relationship can be found between a PCA embedding and listener preference or agreement. In all cases, the linear regression chooses an intercept close to the average of preference ratings or standard deviation of preference and applies a small slope to the independent variable. No R^2 value is greater than 4%. These results disagree with the assessment that "we see

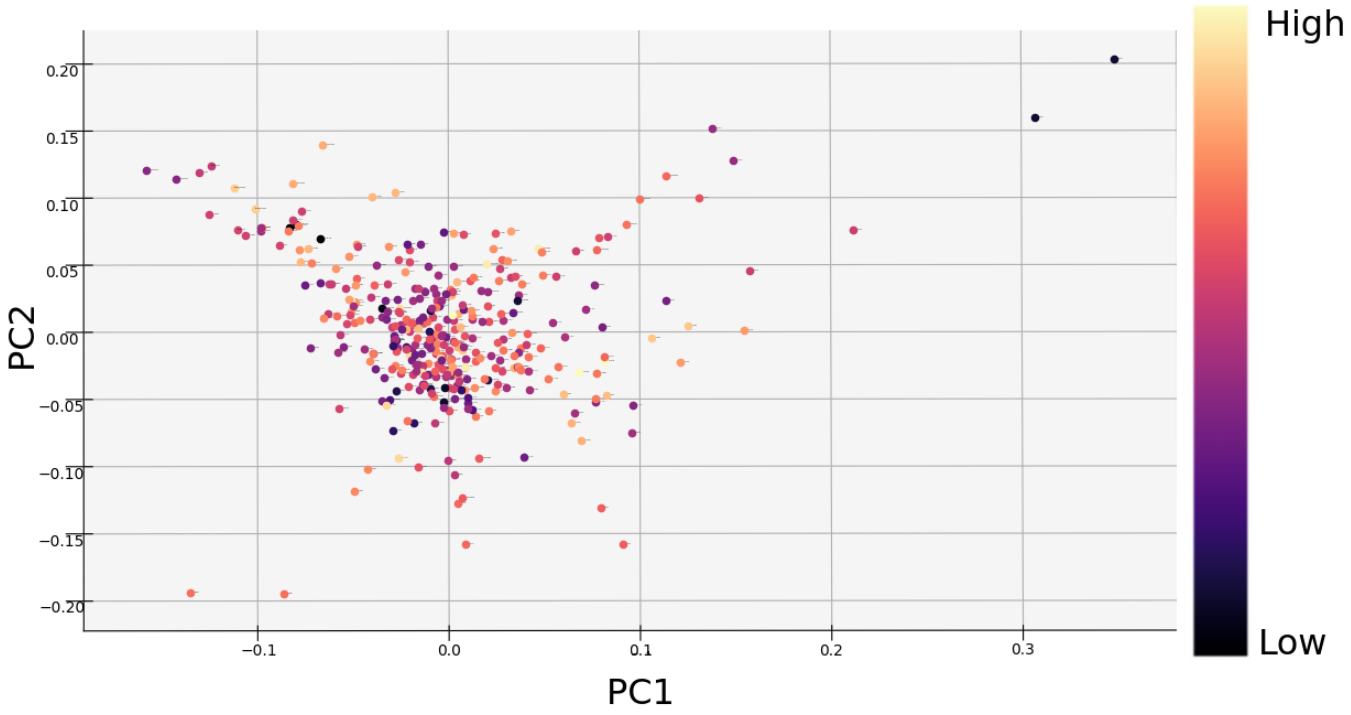

Fig. 3: *Average listener preference for each mix. Results plotted in the first two principal component axes.*

Table 2: *Linear correlation results for average listener preference.*

Dimension	Slope	Intercept	R^2
PC1	0.00915	0.469	1.27×10^{-5}
PC2	-0.0406	0.469	2.26×10^{-4}
PC3	0.0903	0.469	7.62×10^{-4}
$PC1^2$	-1.85	0.475	1.44×10^{-2}
$PC2^2$	0.0246	0.469	8.07×10^{-7}
$PC3^2$	0.0498	0.469	1.56×10^{-5}
Radius	-0.00684	0.470	6.95×10^{-6}

Table 3: *Linear correlation results for standard deviation of listener preference.*

Dimension	Slope	Intercept	R^2
PC1	-0.123	0.197	2.80×10^{-2}
PC2	0.146	0.197	3.56×10^{-2}
PC3	0.0212	0.197	5.11×10^{-4}
$PC1^2$	-0.320	0.198	5.23×10^{-3}
$PC2^2$	-0.389	0.198	2.46×10^{-3}
$PC3^2$	-0.143	0.197	1.55×10^{-3}
Radius	-0.0391	0.199	2.765×10^{-3}

that many higher-quality mixes are located in certain areas of the space [and an] intelligent/automated mixing system could achieve good mixes by "steering" the mix towards these regions " made in [10]. Nor do these results agree with the notion that listener preference would increase with mixes that hit a "sweet spot" in the middle of principal component axes that describe spectral characteristics or loudness. Were higher-quality mixes located centrally within a PCA embedding, one would expect stronger correlations with squared principle component value or radius and average listener preference. Moreover, the lack of correlation with standard deviation of listener preference could suggest that the quality of mixes that strike a balance within the feature space is not agreed upon.

To understand how these principal components relate to perceptual descriptions, the outliers on each axis were cross-referenced with comments made by evaluators in the MED. Mixer DU-J, who took the largest positive value (0.382) in the third principal compo-

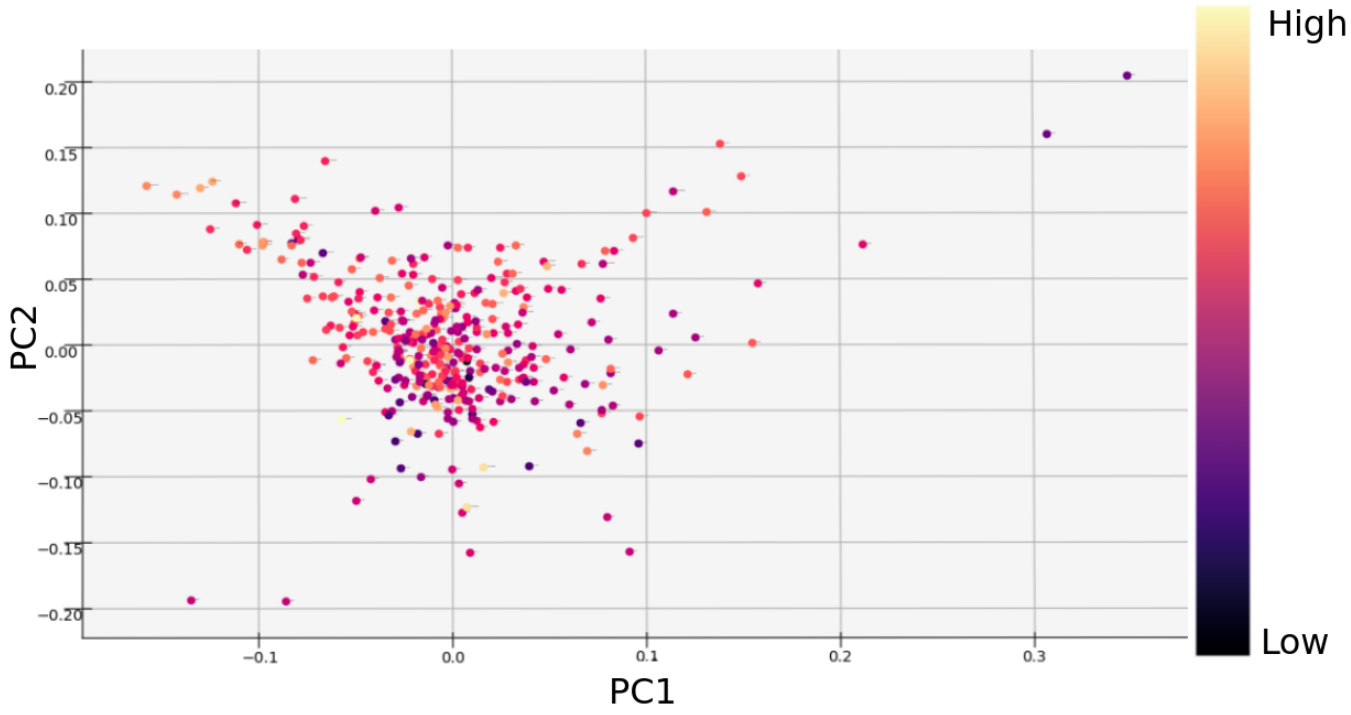

Fig. 4: *Standard deviation of listener preference for each mix. Results plotted in the first two principal component axes.*

nent, received many complaints about the imbalance in its stereo image. Comments included *"it feels like everything comes from the right side"* and *"felt very right-heavy"*. This aligns with the stereo feature loading on the third principal component. However, mixer DU-K, who took a small negative value (-0.020) in the third principal component, also received complaints regarding panning such as *"there must have been some kind of mistake during mixdown [...] the kick is fully panned to the left"*.

While an outlier analysis may reinforce the notion that this principal component analysis can isolate characteristics of a mixdown on each axis, i.e. the stereo image balance maps to the third axis, the full interaction is much more complicated. The weighted summing of 33 features is inherently complex, and while a feature like L/R balance in isolation could point towards extreme panning, the projection to the third axis could be overpowered by features such as EBU Loudness or pmf spread.

Similar explorations of where catch-all terms in the MED are embedded yield similarly complicated results, though a full treatment is outside the scope of this paper. With the current feature set, the PCA embedding concentrates the majority of observations about the origin, making sub-space partitions according to semantic descriptors difficult. Future explorations of this dataset may explicitly consider genre in conjuction with the feature set, or model a specific listener or university group's preferences.

Summary

Principal component analysis was performed on the Mix Evaluation dataset, comprising of 184 mixes across 19 multitrack songs to interrogate listener preference and agreement. 33 features were extracted separately from a verse and chorus for each mix, and projected into a three dimensional principal component embedding space. No strong correlations were found within the PCA embedding space and either average listener preference for a mix or standard deviation of listener preference for a mix. This disagrees with some

claims made in previous studies that suggested certain regions within a PCA embedding contain more highly preferred, or at least agreeable, mixes. Furthermore, text comments made by participants in the MED disagree with previous studies' interpretations of the principal component analysis. Results presented here suggest that the complex nature of loading 33 features onto a single axis prevents PCA axes from fully characterizing distinct aspects of a mix a listener can distinguish.

References

[1] Izhaki, R., *Mixing audio*, Taylor & Francis Group, 2017.

[2] Pras, A. et al., "The impact of technological advances on recording studio practices," *Journal of the American Society for Information Science and Technology*, 64, Mar 2013.

[3] Pestana, P. D., Reiss, J. D., et al., "Intelligent audio production strategies informed by best practices," 2014.

[4] Barcheisi, D. and Reiss, J., "Reverse engineer the mix," *Journal of the Audio Engineering Society*, 58, Jul 2010.

[5] De Man, B., Reiss, J. D., and Stables, R., "Ten years of automatic mixing," in *Proceedings of the 3rd Workshop on Intelligent Music Production*, 2017.

[6] Mimilakis, S.-I., Drossos, K., Floros, A., and Katerelos, D., "Automated tonal balance enhancement for audio mastering applications," in *Audio Engineering Society Convention 134*, 2013.

[7] Mansbridge, S., Finn, S., and Reiss, J. D., "An autonomous system for multitrack stereo pan positioning," in *Audio Engineering Society Convention 133*, Audio Engineering Society, 2012.

[8] Jillings, N. and Stables, R., "Investigating music production using a semantically powered digital audio workstation in the browser," in *2017 AES International Conference on Semantic Audio*, 2017.

[9] Wilson, A. and Fazenda, B., "Variation in multitrack mixes: analysis of low-level audio signal features," *Journal of the Audio Engineering Society*, 64(7/8), pp. 466–473, 2016.

[10] Wilson, A. and Fazenda, B., "101 Mixes: A Statistical Analysis of Mix-Variation in a Dataset of Multitrack Music Mixes," in *139th Convention of the Audio Engineering Society*, NYC, USA, 2015.

[11] De Man, B. and Reiss, J. D., "The mix evaluation dataset," in *20th Int. Conf. on Digital Audio Effects (DAFx-17)*, 2017.

[12] Wilson, A. and Fazenda, B., "Characterization of Distortion Profiles in Relation to Audio Quality," in *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14)*, 2014.

[13] Alluri, V. and Toiviainen, P., "Exploring perceptual and acoustical correlates of polyphonic timbre," *Music Perception: An Interdisciplinary Journal*, 27(3), pp. 223–242, 2010.

[14] EBU-Recommendation, R., "Loudness normalisation and permitted maximum level of audio signals," 2011.

[15] Bogdanov, D. et al., "Essentia: An audio analysis library for music information retrieval," in *14th Conference of the International Society for Music Information Retrieval (ISMIR)*, 2013.

[16] De Man, B. et al., "An Analysis and Evaluation of Audio Features for Multitrack Music Mixtures," in *15th Conference of the International Society for Music Information Retrieval (ISMIR)*, 2014.

[17] Pedregosa, F. et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, pp. 2825–2830, 2011.