Audio Engineering Society

Convention Paper 9848

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA



# Analysis and prediction of the audio feature space when mixing raw recordings into individual stems

Marco A. Martínez Ramírez[1] and Joshua D. Reiss[1]

[1]*Centre for Digital Music, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom*

Correspondence should be addressed to Marco A. Martínez Ramírez (m.a.martinezramirez@qmul.ac.uk)

**ABSTRACT**

Processing individual stems from raw recordings is one of the first steps of multitrack audio mixing. In this work, we explore which set of low-level audio features are sufficient to design a prediction model for this transformation. We extract a large set of audio features from bass, guitar, vocal and keys raw recordings and stems. We show that a procedure based on random forests classifiers can lead us to reduce significantly the number of features and we use the selected audio features to train various multi-output regression models. Thus, we investigate stem processing as a content-based transformation, where the inherent content of raw recordings leads us to predict the change of feature values that occurred within the transformation.

## 1 Introduction

Audio mixing is a crucial part of music production. It essentially tries to solve the problem of unmasking by manipulating the dynamics, spatialisation, timbre or pitch of multitrack recordings [1]. In this paper, we define a *stem* as a processed individual instrument track, and a *raw* track as an unprocessed recording. This differs from subgrouping practices where submixes are created from groups of instruments with similar characteristics [2, 3].

*Stem* processing is an early stage of audio mixing, whose main objective is to combine the *raw* recordings in order to obtain a better representation of the musical source. For example, an electric guitar recorded via different microphone positions plus the direct input is processed into one stereo stem. This manipulation is achieved through a set of linear and nonlinear effects, which can be classified into five different classes: *gain, delay lines, panning, equalisation (EQ) and dynamic range compression (DRC)* [4].

This process is instrument and genre specific since a bass rock *stem* is obtained by applying a different configuration of effects than a bass guitar jazz *stem*. For a specific instrument source this process can be described by (1).

$$s[n] = \sum_{m=1}^{M} H_{m,c}[n] * r_m[n] \tag{1}$$

Where $s$ is the individual processed *stem*, $M$ is the total number of *raw* recordings $r$, $H$ is the chain of audio effects and $c$ their respective control values.

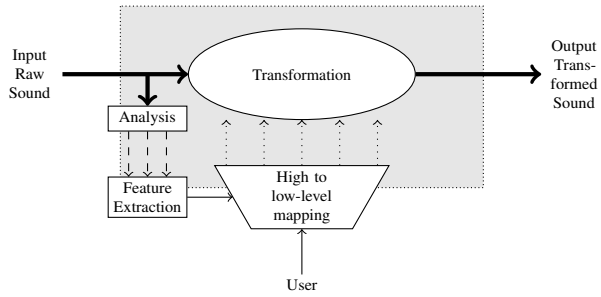



**Fig. 1:** Block diagram of a content-based transformation.

**Table 1:** Most common extracted features

| Feature Type | Feature Name | Frame/Hop size (samples) | Reference |
|---|---|---|---|
| Temporal | log-attack time<br>larm | global<br>2048/1024 | [7]<br>[9] |
| Spectral | spectral centroid<br>Barkbands 1\24<br>spectral contrast coefficients 0\5 | .<br>.<br>. | [7]<br>.<br>[10] |
| Harmonic | odd-to-even ratio<br>hpcp 0\35 | .<br>4096/2048 | [7]<br>[11] |
| Perceptual | specific sharpness<br>specific spread | 2048/1024<br>. | [7]<br>. |

Content-based transformations are described in [5] as the change a particular sound experiences when addressing any type of information related to the audio source, i.e. audio is analysed, meaningful features are extracted and the control signals act to transform the sound and consequently to modify the features. This is similar to how perceptual and high-level features lead the sound engineer through this process. Since the input from the sound engineer is a change of the audio effects' control values, this interaction acts as a high-level transformation that is processed and assigned to the low-level features [5] (see Fig. 1).

Thus, *stem* audio mixing can be modelled as a pre-processing content-based stage where mostly technical criteria is involved, since most of the artistic or creative considerations will take place when blending the *stems* into the final mix. Therefore, we investigate *stem* processing as a content-based transformation, where the inherent content of the *raw* and *stem* tracks can lead us to obtain and predict the affected low-level features.

Our task is to reveal which set of spectral, temporal, harmonic or perceptual low-level features are altered by the transformation in the most consistent way. We explore whether this set of features can be used to design a prediction model. Thus, within a *stem* audio mixing task, we investigate a system that analyses the *raw* recordings and predicts the values of the relevant audio features. Accordingly, the new audio feature values are placed within the *stem* audio feature space and act as an indicator of the expected values for the initial *raw* recordings.

We show that a procedure based on random forests classifiers can lead us to reduce significantly the number of features. We use the selected audio features to train various multi-output regression models. In order to improve the performance, we analyse which set of features are correctly predicted by the models and we explore whether an ensemble of models can provide better predictions.

The rest of the paper is organised as follows. In Section 2 we summarise the relevant literature related to audio feature extraction and feature selection. We formulate our problem in Section 3 and in Section 4 we present the methods. Sections 5, 6 and 7 present the obtained results, their analysis and conclusion respectively.

## 2 Background

### 2.1 Audio features

Extracting audio features is common practice in a variety of fields, such as automatic speech recognition, music information retrieval or audio event recognition. [6] provides a survey of state-of-the-art features in various domains such as temporal, spectral, perceptual, and rhythmic. In a similar way, [7] summarizes a large set of audio features in global and frame-based audio descriptors.

Global features are calculated over the complete audio signal and frame-based or instantaneous features are extracted from overlapping short time windows. The features are retrieved directly from the audio signal or after a respective spectral, harmonic or perceptual transformation. Finally, pooling is performed by modelling the features over time using statistics such as mean, standard deviation, etc. [7, 8].

Some of the most common extracted features can be seen in Table 1.

#### 2.1.1 Audio features and music production

In recent years audio features have been analysed to gain a better understanding of the mixing process or to perform different tasks within an automatic mixing framework. In [12] a wide range of features is extracted from a set of mixing sessions in order to perform an analysis of variance among instruments, songs and sound engineers. [13] explores feature extraction from a set of mixes and it is proposed that higher-quality mixes are located in certain areas of the feature space.

[14] describes sound quality as the lossy compression that the user performs on an audio file. Sound quality classification is achieved by using a selected set of audio features and machine learning classifiers such as SVM and KNN. Feature selection for automatic subgrouping of multitrack audio is performed in [15], where 74 features are selected from a set of 159. Random forests classifiers are used for variable importance and agglomerative clustering for automatic subgrouping.

In [16], *stem* audio mixes were analysed and sets of spectral, dynamic and harmonic audio features were proposed as the main features within a *raw* and *stem* classification task. This is done for different families of instruments such as the bass, guitar, vocals and keys. Furthermore, it was shown that machine learning classifiers improved their performance when using the reduced set of features.

### 2.2 Random forests and feature selection

Random forests classifiers consist of several decision trees that are being constructed and trained using bootstrap aggregation from samples and features of the training data. Bootstrap aggregating, or *bagging*, is a subsampling technique where multiple subsets are drawn at random, but with replacement, from the learning set and consequently used as new learning sets [17]. Therefore, the $k$th decision tree ($t_k$) is trained with a random subset of samples ($l_k$) and each node is split with a random subset of features ($f_k$). In this manner, a random forest classifier consists of a collection of decision trees classifiers $\{clf(\mathbf{x},\Theta_k), k = 1,...\}$ where $\Theta_k$ are independent identically distributed (i.i.d) random vectors containing the subsets $l_k$ and $f_k$. For the input $\mathbf{x}$, the selected class is the mode class among the $k$ tree outputs [18].

Performance is normally measured using the out-of-bag ($OOB$) indicator, which is the average error for each trained tree. It is calculated when $t_k$ predicts the output of a sample that was not included in $l_k$.

Random Forests are also used as indicators of variable importance and two methods are mainly used; the Gini and the permutation importance procedures. The *permutation importance method*, see (2), measures the average decrease of the accuracy on all $OOB$ indicators, when a value of $f_k$ is permuted randomly [19].

$$VI(F_p) = \frac{1}{k}\sum_{t=1}^{k}(OOB_t - OOB_t^p) \tag{2}$$

$VI(F_p)$ is the variable importance of the feature $F_p$, and $OOB_t$ and $OOB_t^p$ are the initial and permuted out-of-bag errors respectively. This method is a more accurate indicator for variable importance and it can be improved when bagging is performed without replacement [20].

## 3 Problem Formulation

For a specific instrument source, consider $M$ *raw* recordings $r$ and one processed *stem* $s$, for which we extract and pool a set of audio features $F^r$ and $F^s$ respectively. We use a procedure based on random forests classifiers ($rfc$) and the permutation method ($VI$) to reduce the number of audio features.

We attempt to find the set of features ($f_{pred}$) in order to build a prediction model of the transformation.

$$rfc\{r,s\}, VI\{F^r, F^s\} \Longrightarrow f_{pred} \tag{3}$$

Thus, given the input vector of *raw* prediction features values $r\{f_{pred}\} = f_1^r, f_2^r, ..., f_n^r$ and the target vector of *stem* prediction features values $s\{f_{pred}\} = f_1^s, f_2^s, ..., f_n^s$, where $n$ is the number or prediction features $|f_{pred}|$. We train different regression models to learn the following function.

$$y(\cdot) : r\{f_{pred}\} \rightarrow s\{f_{pred}\} \tag{4}$$

Finally, using regression metrics, we explore which subset of features and which combination of models lead to the best predictions.

# 4 Methods

## 4.1 Dataset

The *raw* recordings and individual processed *stems* were taken from [21], mostly based on [22] and following the same structure; a song consists of the mix, *stems* and *raw* audio. 102 multitracks were selected which correspond to genres of commercial western music such as *Rock, Folk, Jazz, Pop, Fusion* and *Rap*. These have been mixed by experienced sound engineers and recorded in professional studios.

All tracks have a sampling frequency of 44.1 kHz, and we proceeded to find the 10 seconds with the highest energy for each *stem* track. Our assumption is that the most relevant *raw* recording is the one with the highest energy. Thus, the corresponding *raw* tracks were analysed during the same 10 second interval and the one with the highest energy was chosen. The selected tracks were downmixed to mono, loudness normalisation was performed using *replayGain* and an equal-loudness filter [23]. The test dataset corresponded to 10% of the *raw* and *stem* tracks. Table 2 shows the dataset.

**Table 2:** Raw and stem number of tracks by instrument group.

| Group | Instrument Source | Raw | Stem |
|---|---|---|---|
| bass | electric bass | 96 | 62 |
| | synth bass | 12 | 6 |
| guitar | clean electric guitar | 112 | 36 |
| | acoustic guitar | 55 | 24 |
| | distorted electric guitar | 78 | 20 |
| | banjo | 2 | 2 |
| vocal | male singer | 145 | 36 |
| | female singer | 61 | 22 |
| | male rapper | 12 | 2 |
| keys | piano | 113 | 38 |
| | synth lead | 51 | 17 |
| | tack piano | 27 | 7 |
| | electric piano | 3 | 3 |

## 4.2 Feature Extraction

Based on the evaluation of audio feature extraction libraries presented in [24], the spectral, temporal, harmonic and perceptual low-level features were extracted using [25]. In total, 78 different features were extracted, of which 15 are global and 63 are frame-based descriptors. Most of the frame-based features were computed with frame/hop sizes equal to 2048/1024 samples, although there were some exceptions with sizes of 4096/2048 and 88200/44100 samples.

Pooling was performed over the frame-based features and the following statistics were calculated: *mean, median, variance, standard deviation, minimum, maximum, kurtosis, skewness and mean and variance of the first and second derivatives*. Thus, from each *stem* and *raw* segment, a total of $|F| = 1812$ features were extracted.

## 4.3 Feature Selection

In order to perform the selection of features, we followed the same procedure as [16], which is based on [19]. The following steps allowed us to obtain the prediction features $f_{pred}$.

### 4.3.1 Prediction features

- A total of 50 random forests classifiers with $k = 2000$ and $|f_k| = |F|/3$ were built.
- The mean of the feature importances along with their corresponding standard deviations were sorted in descending order. Feature importance was calculated with (2).
- The threshold of importance was estimated by fitting the standard deviation values with a decision tree regressor and retaining only the features with importance value above this threshold. These are the first preselected features $f_{p1}$.
- A nested set of random forest classifiers were constructed using the preselected features. This was done starting from the most important feature and one feature was added for each classifier that was built. All classifiers were fitted 50 times and two labels were used in the classification task: *raw* and *stem*. We selected the features that led to the minimum mean *OOB* error. These are the second preselected features $f_{p2}$.
- Using $f_{p2}$ an ascending sequence of random forests classifiers is built following the same principles as the nested set. The difference is that a feature is only added if the decrease of the *OOB* error is significant. This threshold is defined by (5). It is the mean of the absolute value of the first derivative of the *OOB* errors, corresponding to the models trained with the set of features $(f_{p1} \cap f_{p2})^c$.

- Each random forest classifier is fitted 50 times, and the features of the last model correspond to the prediction features $f_{pred}$.

$$TH_{pred} = \frac{1}{|f_{p1}| - |f_{p2}|} \sum_{j=|f_{p2}|}^{|f_{p2}|-1} |OOB(j+1) - OOB(j)| \quad (5)$$

### 4.4 Feature regression models

We performed a 5-fold cross-validation to optimise the hyperparameters of different multi-output regression models. Hyperparameters are the parameters whose values are set before the training process. The trained models are: support vector regressor (SVR), random forest regressor (RF), k-nearest neighbour (KNN), partial least squares (PLS) and linear regression (LR). This was done for each instrument group and its $f_{pred}$ set. The optimal hyperparameters are presented in Table 3. Each KNN regressor was trained using the *minkowski* distance as metric.

The models were evaluated with the mean average percentage error (*mape*) and the mean absolute error (*mae*).

$$mape(f) = \frac{1}{N} \sum_{i=0}^{N} \frac{|y^i_{test} - y^i|}{y^i_{test}} \quad (6)$$

$$mae(f) = \frac{1}{N} \sum_{i=0}^{N} |y^i_{test} - y^i| \quad (7)$$

Where $y$ and $y_{test}$ are the predicted and real *stem* values of a specific feature $f \in f_{pred}$. $N$ is the number of *stem* segments in the test dataset.

An entry was considered an outlier if its value was greater than 3 times the standard deviation after subtracting the mean. These were removed from the training dataset.

**Table 3:** Hyperparameters of the multi-output regression models.

| Group | SVR | | | | RF | KNN | PLS |
|---|---|---|---|---|---|---|---|
| | kernel | C | gamma | epsilon | trees ($k$) | n | components |
| bass | rbf | 1 | 8 | 0.002 | 1000 | 6 | 1 |
| guitar | rbf | 1024 | 3.12e-2 | $0.14\overline{9}$ | 750 | 19 | 1 |
| vocal | sigmoid | 64 | 1e-4 | 0.73 | 1000 | 13 | 1 |
| keys | sigmoid | 32768 | 1.25e-4 | $0.29\overline{9}$ | 2000 | 5 | 4 |

## 5 Results

The feature selection procedure was applied to the bass, guitar, vocal and keys instrument groups.

First, Fig. 2 shows the mean of importance in descending order for the first 50 features. Then, based on the standard deviation of the importances, the threshold estimation led to the first preselected features ($f_{p1}$) and from the nested set of random forests classifiers $f_{p2}$ was established. Finally, $f_{pred}$ was obtained by constructing an ascending set of random forest classifiers whose *OOB* error is shown in Fig. 3.

The number of obtained features and the list of prediction features is presented in Table 4 and Table 5 respectively.

The multi-output regression models were trained with the *raw* and *stem* $f_{pred}$ vectors as input and target respectively. The performance of the models can be seen in Table 6.

The features that led to the best performance were used to re-train the models. Using these features we obtained the best possible ensemble of regression models. Accordingly, the average *mape* (*a-mape*) between the selected features was used as a metric. This was done for each group of instruments and the results can be seen in Table 7.

In order to have a visual cue from the latter set of features, we used PCA to fit the *raw* and *stem* feature vectors into a two-dimensional space. This can be seen in Fig. 4.

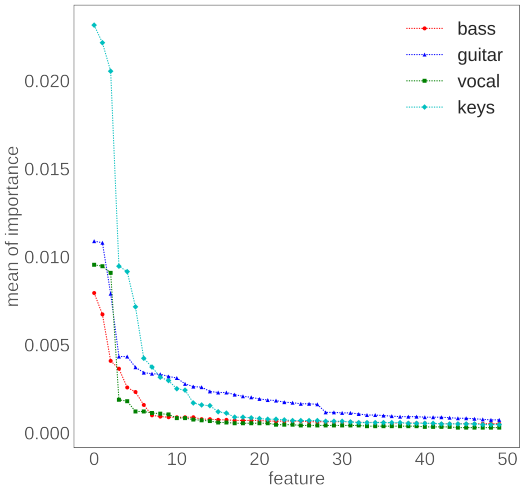



**Fig. 2:** Mean of importance for the first 50 features for bass, guitar, vocal and keys.

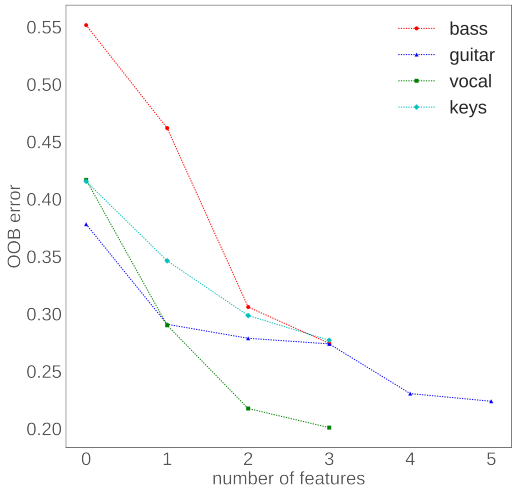



**Fig. 3:** OOB error and number of features for the ascending set of random forest classifiers.

**Table 4:** Number of preselected and prediction features.

| Group | $\lvert f_{p1} \rvert$ | $\lvert f_{p2} \rvert$ | $\lvert f_{pred} \rvert$ |
|---|---|---|---|
| bass | 7 | 6 | 4 |
| guitar | 28 | 7 | 6 |
| vocal | 14 | 7 | 4 |
| keys | 24 | 14 | 4 |

## 6 Analysis

### 6.1 Feature selection

Fig. 2 shows that from 1812 features no more than 30 have a significant mean of importance. $f_{p1}$ is larger for keys and guitar ($> 20$) than for bass and vocal ($< 15$). When selecting $f_{p2}$ the feature set size is further reduced, having 14 features for keys and less than 7 features for bass, guitar and vocal. The size of $f_{pred}$ was fairly uniform between the groups, with 4 features for bass, vocal and keys and 6 features for guitar.

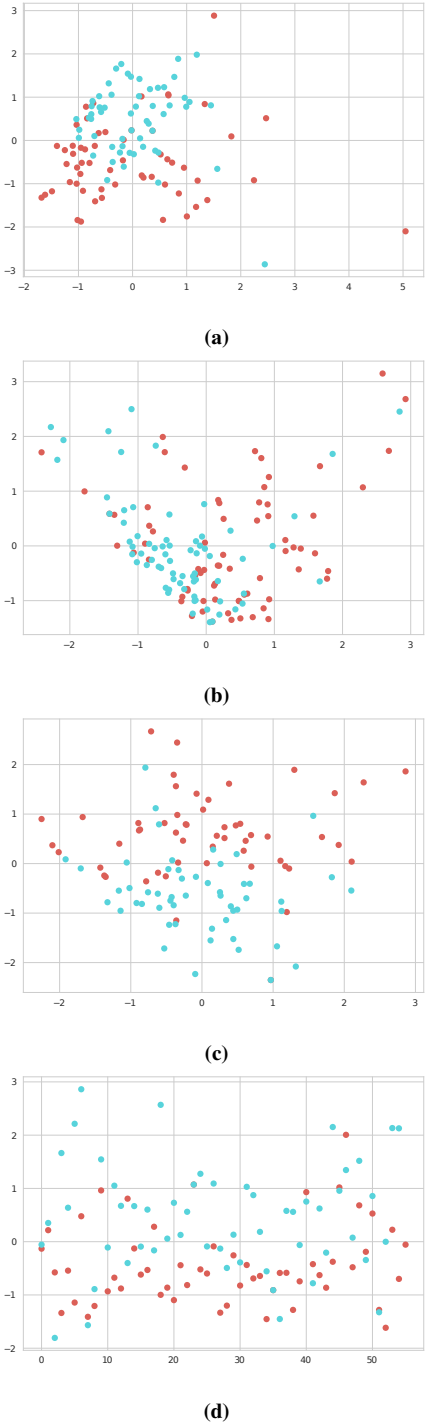
From Fig. 3, it can be seen that the order of $f_{pred}$ is based on features that reduce the most the *OOB* error. Table 5 shows that the selected features correspond to different types of dynamic, spectral and harmonic audio characteristics.

### 6.2 Features for the prediction of the transformation

From Table 6 it is evident that the performance of the models was satisfactory for certain features. *Mape* values of less than 10% were observed for features belonging to the bass, guitar, and vocal groups. On the other hand, 8 features across all instruments' groups presented deficient results, with *mape* values greater than 50% and in some cases extremely large values. Also, the *mae* helps to understand the relative *mape* in terms of the units of measurement of the respective features.

In general, when considering the individual features, LR provided the best results. SVR, KNN and PLS performed consistently as well as RF. The latter was more robust to the features in which the other models failed, but still achieved poor results (*mape* $> 80\%$).

The multi-output regression models were optimised by removing the features the models failed to provide a good estimate for. Table 7 shows that a greater number of characteristics remained for the vocal and bass groups. The best results were obtained from an ensemble composed of SVR and LR. Both bass and vocal, achieved *a-mape* values around 12%. Within the guitar and keys groups, the models predicted a smaller



**(a)** **(b)** **(c)** **(d)**

**Fig. 4:** PCA visualisation for (a) bass, (b) guitar, (c) vocal and (d) keys. Raw and stem segments are red and blue respectively. Axes are unitless.

**Table 5:** List of prediction features.

| Group | Name | Pooling |
|---|---|---|
| bass | 1 - spectral contrast valley (0)<br>2 - effective duration<br>3 - hpcp (33)<br>4 - spectral energy low | max<br>global<br>variance second derivative<br>mean |
| guitar | 1 - rms<br>2 - spectral energy middle-low<br>3 - loudness stevens<br>4 - spectral energy middle-low<br>5 - spectral contrast valley (0)<br>6 - loudness stevens | variance first derivative<br>variance second derivative<br>variance second derivative<br>mean first derivative<br>max<br>mean second derivative |
| vocal | 1 - larm<br>2 - spectral contrast coeff. (1)<br>3 - pitch salience<br>4 - pitch salience | variance first derivative<br>standard deviation<br>mean first derivative<br>mean second derivative |
| keys | 1 - spectral energy middle-low<br>2 - spectral energy<br>3 - loudness vickers<br>4 - barkbands (4) | variance<br>max<br>max<br>max |

number of features correctly. SVR and RF provided *a-mape* values around 20%, which is still considered acceptable.

This is also confirmed by Fig. 4. It is easier to differentiate between the *raw* and *stem* segments for the vocal and bass groups, while guitar and keys are more difficult to separate.

The optimal features for the prediction of the transformation can be classified into the following classes.

#### 6.2.1 Temporal / Dynamic

Dynamic features associated with loudness were present for the guitar, vocal and keys groups. These are related to the *long-term loudness (larm)* [9], *loudness stevens* [26] and *loudness vickers* [27]. For the bass, *effective duration* [7] was present, which is a global temporal indicator associated to the envelope of an audio segment. Since loudness normalisation was applied prior to the extraction of the features, this transformation of feature values could be an indicator of the rate of change of loudness, or the modification of the envelope due to DRC.

**Table 6:** Evaluation of the multi-output regression models. The bold entries represent the minimum error values.

| Group | $f_{pred}$ | SVR | | RF | | KNN | | LR | | PLS | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *mape* | *mae* | *mape* | *mae* | *mape* | *mae* | *mape* | *mae* | *mape* | *mae* |
| bass | 1 | 0.119 | 0.7403 | 0.1084 | 0.6523 | 0.1113 | 0.6763 | 0.1065 | 0.6441 | **0.0991** | **0.5994** |
| | 2 | 0.0055 | 0.0532 | **0.0054** | **0.0492** | 0.0071 | 0.0696 | 0.0119 | 0.1168 | 0.0087 | 0.0852 |
| | 3 | 0.3659 | 0.0249 | 0.5068 | 0.035 | 0.4444 | 0.0299 | **0.2652** | **0.0165** | 0.351 | 0.0222 |
| | 4 | 33.746 | 6.6e-4 | 7.6611 | 3.2e-4 | **4.6765** | **2.9e-4** | 11.2052 | 2.9e-4 | 11.996 | 3.3e-4 |
| guitar | 1 | 6.3777 | 2.1e-5 | **0.8759** | 8.8e-6 | 1.7194 | 8.4e-6 | 1.8354 | **7.6e-6** | 1.7958 | 8.9e-6 |
| | 2 | 597 | 3.9e-5 | **2.0051** | **1.4e-6** | 23.82 | 2.7e-6 | 14.88 | 2.2e-6 | 27.50 | 3.0e-6 |
| | 3 | 2.6341 | 0.1409 | **1.0023** | 0.1480 | 3.0576 | **0.1295** | 3.846 | 0.1319 | 3.4238 | 0.1416 |
| | 4 | 16.678 | 2.4e-3 | **0.5617** | 2.3e-3 | 1.6513 | 3.0e-4 | 1.1104 | **2.6e-4** | 1.8755 | 3.4e-4 |
| | 5 | 0.1105 | 0.9320 | 0.1172 | 0.9861 | 0.1283 | 1.1126 | **0.098** | **0.8108** | 0.1124 | 0.9305 |
| | 6 | **0.3741** | **0.1541** | 0.4448 | 0.2170 | 0.4273 | 0.1625 | 0.5606 | 0.1765 | 0.4978 | 0.1743 |
| vocal | 1 | 0.2774 | 2.0896 | 0.3508 | 2.5153 | 0.2721 | 2.0062 | **0.2697** | **1.7983** | 0.2758 | 2.0146 |
| | 2 | 0.1078 | 0.0145 | 0.0783 | 0.0112 | 0.0769 | 0.0111 | 0.0934 | 0.0132 | **0.071** | **0.0105** |
| | 3 | 0.1538 | 0.0111 | 0.1442 | 0.0102 | 0.2003 | 0.0144 | **0.0977** | **0.0069** | 0.1806 | 0.0129 |
| | 4 | 0.1579 | 0.0178 | 0.1499 | 0.0171 | 0.2116 | 0.0242 | **0.0991** | **0.0111** | 0.1872 | 0.0213 |
| keys | 1 | 3241 | 2.5e-5 | **81.91** | 6.6e-6 | 346.5 | 7.1e-6 | 405.5 | 7.4e-6 | 315.8 | 6.8e-6 |
| | 2 | 10.70 | 0.0148 | **2.42** | **0.0088** | 5.16 | 0.0104 | 3.98 | 0.0102 | 4.05 | 0.0095 |
| | 3 | 0.1924 | 5.5232 | **0.1911** | 5.4044 | 0.2195 | 6.2929 | 0.1944 | **5.3863** | 0.2033 | 5.7686 |
| | 4 | 22.35 | 6.8e-3 | 5.47 | 4.8e-3 | 11.22 | 5.2e-3 | **4.16** | **4.5e-3** | 7.851 | 5.0e-3 |

**Table 7:** Average absolute relative error of the multi-output regression models that performed the best with the optimal set of prediction features.

| Group | $f_{pred}$ | models | *a-mape* |
|---|---|---|---|
| bass | 1-2-3 | LR | 0.1279 |
| guitar | 5-6 | SVR | 0.2423 |
| vocal | 1-2-3-4 | SVR-LR | 0.1263 |
| keys | 3 | RF | 0.1911 |

#### 6.2.2 Spectral

The second *spectral contrast coefficient* and the first *spectral contrast valleys*, which are related to the shape of the spectrum [28], are among the selected features for the bass, guitar and vocal groups. This could be an indicator of common practices or a targeted contour when applying EQ to these types of instruments.

On the other hand, the features that the model failed to predict are mainly related to spectral energy measurements: *middle-low spectral energy* (150Hz-800Hz) for guitar and keys, *total spectral energy* and the fourth *barkband* (300Hz) also for the keys and the mean *low spectral energy* (20Hz-150Hz) for the bass. These frequency bands are as expected since they contain most of the energy of the respective instruments [29]. Another common factor is that most units of measurement of these characteristics correspond to values close to zero ($< 1e-4$).

#### 6.2.3 Harmonic

For the bass, the 33rd *harmonic pitch class profile (HPCP)* was one of the selected features. The *HPCP* is calculated from the spectral peaks and represents the intensities of various subdivisions of semitone pitch classes [11]. This feature is expected to be related to the harmonic distortion due to the application of overdrive audio effects, which are used to enhance certain harmonics or to reinforce some sounds within the mix [30]. For the vocal, the harmonic features are associated to the *pitch salience*, which is a measure of the tone sensation linked to the autocorrelation of the signal [31]. This feature is prone to be related to the application of pitch shifting correction or de-essing [30] to the vocal tracks.

## 7 Conclusion

In this work, we determined the sets of audio features that can be used to predict *stem* audio mixing as a content-based transformation. We extracted a set of 1812 audio features from bass, guitar, vocal and keys *raw* recordings and *stems*. We used a procedure based on random forest classifiers to find the features that were altered the most consistently by the transformation. We trained various multi-output regression models and based on their performances, we further reduced the number of features that can be used to predict the transformation correctly.

The dynamic, spectral and harmonic feature values of the vocal and bass *stem* segments were correctly predicted from the respective *raw* recording feature values. The same was achieved for the keys and guitar dynamic and spectral feature values, although with a smaller set of features and a larger margin of error.

Since the keys and guitar groups were composed of a more diverse range of instruments, this might be a reason for the reduced generalisation by the regression models. Also because these instruments often have diverse roles among the different genres. On the other hand, vocal and bass can be considered to be mixed in a more consistent and regular way. Therefore, the regression models learned and predicted the transformation more accurately.

We conclude that the underlying characteristics of manipulating *raw* recordings into individual *stems* can be described by the mapping of a selected set of audio features. Thus, we provide a framework to guide automatic mixing systems or sound engineers within *stem* mixing tasks.

As a future work, the preprocessing of features that models failed to predict could improve the performance of regressors, e.g. a way to meaningfully scale these feature values. Also, Linear Regression was one of the models that better performed and this indicates a linear relationship among the transformation. Nevertheless, further analysis of the feature values transformation is required, in addition to how these relate to specific audio effects. Finally, the method can be improved by improving the selection of *raw* recordings, so that more than one is taken into account during feature extraction.

## References

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

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

[3] Ronan, D. et al., "The impact of subgrouping practices on the perception of multitrack mixes," in *139th Audio Engineering Society Convention*, 2015.

[4] Pestana, P. D. and Reiss, J. D., "Intelligent audio production strategies informed by best practices," in *53rd Conference on Semantic Audio: Audio Engineering Society*, 2014.

[5] Amatriain, X. et al., "Content-based transformations," *Journal of New Music Research*, 32(1), pp. 95–114, 2003.

[6] Mitrović, D., Zeppelzauer, M., and Breiteneder, C., "Features for content-based audio retrieval," *Advances in computers*, 78, pp. 71–150, 2010.

[7] Peeters, G., "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," 2004.

[8] Hamel, P. et al., "Temporal Pooling and Multiscale Learning for Automatic Annotation and Ranking of Music Audio." in *12th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 729–734, 2011.

[9] Nielsen, S. H. and Skovenborg, E., "Evaluation of different loudness models with music and speech material," in *117th Audio Engineering Society Convention*, 2004.

[10] Jiang, D.-N. et al., "Music type classification by spectral contrast feature," in *IEEE International Conference on Multimedia and Expo (ICME)*, volume 1, pp. 113–116, 2002.

[11] Gómez, E., "Tonal description of polyphonic audio for music content processing," *INFORMS Journal on Computing*, 18(3), pp. 294–304, ????

[12] De Man, B. et al., "An analysis and evaluation of audio features for multitrack music mixtures," in *15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[13] 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.

[14] Fourer, D. and Peeters, G., "Objective characterization of audio signal quality: applications to music collection description," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[15] Ronan, D. et al., "Automatic subgrouping of multitrack audio," in *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15)*, 2015.

[16] Martínez Ramírez, M. A. and Reiss, J. D., "Stem audio mixing as a content-based transformation of audio features," in *19th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2017.

[17] Breiman, L., "Bagging predictors," *Machine learning*, 24(2), pp. 123–140, 1996.

[18] Breiman, L., "Random forests," *Machine learning*, 45(1), pp. 5–32, 2001.

[19] Genuer, R., Poggi, J.-M., and Tuleau-Malot, C., "Variable selection using random forests," *Pattern Recognition Letters*, 31(14), pp. 2225–2236, 2010.

[20] Strobl, C. et al., "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, 8(1), p. 25, 2007.

[21] De Man, B. et al., "The open multitrack testbed," in *137th Audio Engineering Society Convention*, 2014.

[22] Bittner, R. M. et al., "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research." in *15th International Society for Music Information Retrieval Conference (ISMIR)*, volume 14, pp. 155–160, 2014.

[23] Wolters, M., Mundt, H., and Riedmiller, J., "Loudness normalization in the age of portable media players," in *128th Audio Engineering Society Convention*, 2010.

[24] Moffat, D. et al., "An evaluation of audio feature extraction toolboxes," in *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15)*, 2015.

[25] Bogdanov, D. et al., "Essentia: An Audio Analysis Library for Music Information Retrieval." in *14th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 493–498, 2013.

[26] Teghtsoonian, R., Stevens, S., and Stevens, G., "Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects," *The American Journal of Psychology*, 88(4), p. 677, 1975.

[27] Vickers, E., "Automatic long-term loudness and dynamics matching," in *111th Audio Engineering Society Convention*, 2001.

[28] Akkermans, V., Serrà, J., and Herrera, P., "Shape-based spectral contrast descriptor," in *Proceedings of the 6th Sound and Music Computing Conference (SMC)*, pp. 143–148, 2009.

[29] Giordano, N., "Spectral Analysis of Musical Sounds with Emphasis on the Piano," *The Journal of the Acoustical Society of America*, 138(2), pp. 846–846, 2015.

[30] Reiss, J. D. and McPherson, A., *Audio effects: theory, implementation and application*, CRC Press, 2014.

[31] Terhardt, E., Stoll, G., and Seewann, M., "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *The Journal of the Acoustical Society of America*, 71(3), pp. 679–688, 1982.