# Stem Audio Mixing as a Content-Based Transformation of Audio Features

Marco A. Martínez Ramírez, Joshua D. Reiss
Centre for Digital Music
Queen Mary University of London
{m.a.martinezramirez,joshua.reiss}@qmul.ac.uk

***Abstract*—Multitrack audio mixing is an essential part of music production and one of the first steps consist on processing individual stems from raw recordings. In this paper, we investigate this stage as a content-based transformation. We explore which audio features are relevant to interpret this specific process and which set of features gets modified by the mixing of stems in the most consistent way. We show that the number of features can be reduced with a procedure based on the permutation importance method of random forest classifiers. Thus, the selected audio features are used to train various classification models and we analyse which set of features lead to a better classification accuracy. We conclude that the underlying characteristics of manipulating raw recordings into individual stems can be described by this selected set of features.**

## I. Introduction

AUDIO MIXING is a highly cross-adaptive transformation since the processing of an individual track depends on the content of all tracks involved [1]. This transformation is performed through a set of linear and nonlinear effects which can be classified into five classes: *gain, delay, panning, equalisation (EQ) and dynamic range compression (DRC)* [2].

We define a *stem* as a processed individual instrument track, and a *raw* track as an unprocessed recording. This differs from subgrouping practices where the mixing engineer groups instruments into submixes in order to manipulate a large number of tracks at once [3], [4].

Stem audio mixing is the processing of various raw tracks into an individual stem. Each corresponding to a distinct instrument or sound source; i.e. a guitar recorded via different microphone positions is processed into one stereo stem.

The main goal of this step is to process the individual source tracks separately prior to blend them into a final mix. In this manner, this transformation can be seen as part of multi-microphone signal processing, where the task is to combine the available recordings in order to obtain a better representation of the musical source. For a specific instrument source this process can be described by Fig. I and (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.

978-1-5090-3649-3/17/$31.00 ©2017 IEEE

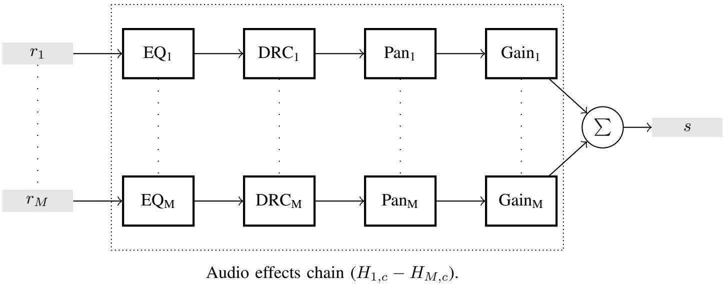


Fig. 1. Block diagram of the transformation of raw recordings into stems.

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. Such type of processing is also proposed by [6] as adaptive audio effects.

Thus, in this work we investigate stem processing as a content-based transformation, where we explore which set of low-level audio features change in the most consistent way. We use the selected audio features to train various classification models and we analyse which set of features leads to a better classification between raw and stem tracks. We investigate whether these features can inform us about the fundamental audio characteristics that sound engineers manipulate when performing this step. In related work [7], we build on these results by using the selected audio features to train various multi-output regression models.

This paper is organised as follows. In Section II we present the relevant literature. We formulate our problem and experiment in Section III and IV respectively. Sections V and VI show the results and their analysis. We conclude with Section VII.

## II. Background

### A. Audio Features

A survey of state-of-the-art audio features is presented in [8]. In a similar way, [9] 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 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 them over time using statistics such as mean, standard deviation, etc., [9], [10].

Also, audio features have been analysed for automatic mixing tasks or to gain a better understanding of the mixing process. [11] evaluated audio feature variance among instruments, songs and sound engineers. Similarly, [12] proposed that higher quality mixes and certain values of audio features are related.

Sound quality classification of individual tracks was performed in [13]. This, by using a selected set of audio features and through different machine learning classifiers. Feature selection is achieved through random forest classifiers in [14], where the selected descriptors are used for automatic subgrouping of multitrack audio.

To the best of our knowledge, feature selection has not been implemented for individual stem processing from raw recordings.

### *B. Random Forests and Variable Importance*

Random Forests is an ensemble learning method for both classification and regression problems. It consists of several decision trees that are being constructed and trained using bootstrap aggregating from samples and features of the learning 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 [15]. 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 ($f_k$) from the complete learning set ($L$) and feature set ($\bar{F}$) respectively. 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 **x**, the selected class is the mode class among the *k* tree outputs [16].

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 Gini method provides a ranking of the variables that is related to the mean entropy loss in each split node when growing trees with different subsets of $f_k$. This method is much faster to calculate, although it is more biased, more unstable and is not robust to the variation of units of measure or the number of categories among all variables [17]. 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 [18].

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

$VI(F_p)$ 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 [17]. None of these methods are robust when estimating the variable importance of highly correlated variables [18].

## III. 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 model stem audio mixing as a content-based transformation of audio features:

$$\sum_{M} r\{F^r\} \longmapsto s\{F^s\} \tag{3}$$

We use a procedure based on Random Forests classifiers ($clf$) and the permutation variable importance method ($VI$) to reduce the number of audio features. We attempt to find two subsets of features: *1)* important features for interpretation of the transformation ($f_{int}$), *2)* a small number of features to build a prediction model of the transformation ($f_{pred}$).

$$VI\{clf(F^r, F^s)\} \Longrightarrow f_{int}, f_{pred} \tag{4}$$

The $clf$ are trained with $F^r$ and $F^s$ as input vectors and with the *raw* and *stem* labels as the output classes. Thus, $VI$ is used over the trained classifiers to obtaining the feature subsets that get modified by the mixing of stems in the most consistent way.

Finally, different machine learning classifiers are trained with $F$, $f_{int}$, and $f_{pred}$ and we explore which subset of features leads to a better classification accuracy.

## IV. Experiment

### *A. Dataset*

The raw recordings and individual processed stems were taken from [19], mostly based on [20] and following; 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. Table I shows the dataset.

### *B. Feature Extraction*

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 then analysed and the one with the highest energy in the same 10 second interval was chosen. We decided this was the best generalisation since there are currently no proposals or established available rules on how to mix raw recordings in order to obtain stem tracks.

The selected segments were downmixed to mono and loudness normalisation was performed using *replayGain* and

TABLE I
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 |

an equal-loudness filter [21]. All the low-level descriptors available in [22] were extracted. 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, a total of 1812 features $|F|$ were extracted from each stem and raw segment.

### C. Feature Selection

In order to perform the selection of features, the procedure proposed in [18] was followed. The following steps allowed us to obtain $f_{int}$ and $f_{pred}$.

*1) Interpretation 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 preselected features $f_p$.

- A nested set of random forest classifiers was constructed with 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 interpretation features $f_{int}$.

*2) Prediction features:*

- An ascending sequence of random forests classifiers was built, only that this time 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 $(f_p \cap f_{int})^{\mathsf{c}}$.

TABLE II
PARAMETERS OF THE CLASSIFIERS.

| RF | | SVM | | | LG |
|---|---|---|---|---|---|
| trees $(k)$ | $\lvert fk \rvert$ | kernel | C | gamma | C |
| 2000 | $\lvert\bar{F}\rvert/3$ | rbf | 1 | $1/\lvert\bar{F}\rvert$ | 1 |

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

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

*3) Raw and Stem classifiers:*

Random Forests (RF), Support Vector Machine (SVM) and Logistic Regression (LG) classifiers were trained with $F$, $f_{int}$, and $f_{pred}$. This was done using a test subset, which corresponds to 10% of the original dataset and it was not used in the feature selection process. Table II shows the parameters for each classifier.

## V. 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 for all four instrument types. Fig. 3 shows, for the sake of clarity, only the estimated threshold and the decision tree regression curve for the Vocal's standard deviation of importance.

The threshold estimation leads to a set of 7, 28, 14 and 24 preselected features ($|f_p|$) for Bass, Guitar, Vocal and Keys respectively. In order to obtain $f_{int}$, the nested set of random forest classifiers was constructed using $f_p$ and the $OOB$ error is shown in Fig. 4. Likewise, $f_{pred}$ was obtained by constructing an ascending set of random forest classifiers whose $OOB$ error is shown in Fig. 5. The list of interpretation and prediction features is presented in Table III and IV.

In addition, a heatmap of correlation among the $f_{pred}$ of each group of instruments is presented in Fig. 6. Finally, the results of the different machine learning classifiers are shown in Table V.

## VI. ANALYSIS AND DISCUSSION

Fig. 2 shows that from 1812 features no more than 30 have a significant mean of importance. $f_p$ is larger for Keys and Guitar ($> 20$) than for Bass and Vocal ($< 15$). When selecting $f_{int}$ the feature set size is further reduced, having 14 features for the Keys and less than 7 features for Bass, Guitar and Vocal. This is because the Keys group has more variation in the instruments that compose it, since it contains a more diverse type of sound sources. The size of $f_{pred}$ was fairly uniform with 6 or fewer features across each instrument groups.

From Table III and IV, the order of the features $f_{int}$ is aligned with the mean importance descending order obtained from the permutation method. Whereas the $f_{pred}$ order of

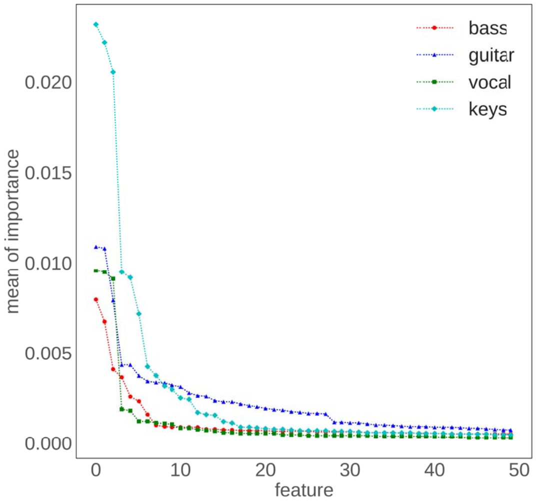


Fig. 2. Mean of importance for the first 50 ranked features for Bass, Guitar, Vocal and Keys.

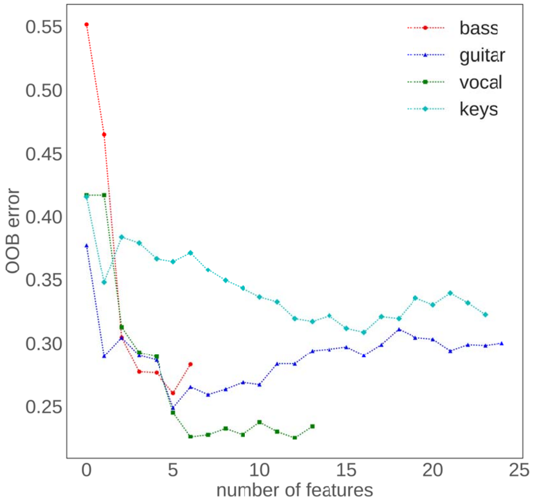


Fig. 4. OOB error and number of features for the nested set of random forests classifiers.

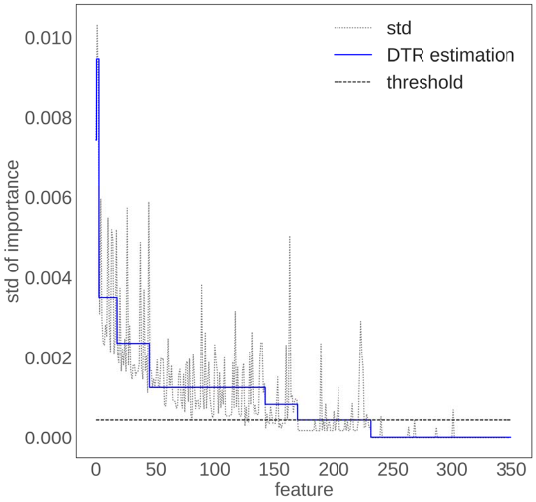


Fig. 3. Standard deviation of importance, decision tree regression (DTR) curve and estimated threshold of importance for the first 350 features for the Vocal group.

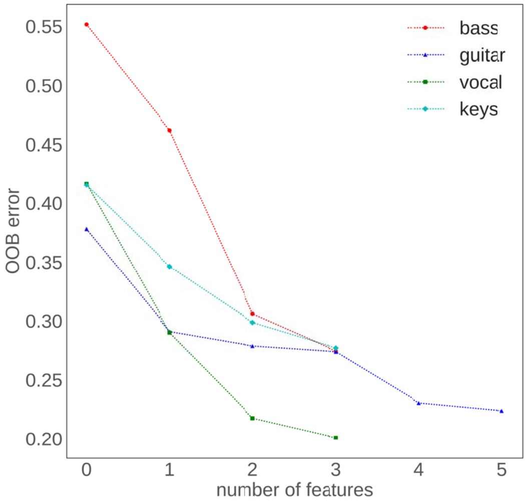


Fig. 5. OOB error and number of features for the ascending set of random forest classifiers.

the features is based on features that reduce the most the $OOB$ error. For this reason, by reducing $f_{int}$ into $f_{pred}$, the procedure leads to a less biased order of features.

### A. Prediction features

The majority of the set of features are associated to the energy and the shape of the spectrum. *Middle-low spectral energy* (150Hz-800Hz) measurements are present for the Guitar and Keys, in addition to the *total spectral energy* and the fourth *barkband* (300Hz) for the Keys. Also, the mean *low spectral energy* (20Hz-150Hz) is among the prediction features for the Bass. These frequency bands are as expected since they contain most of the energy of the respective instruments [23]. The *spectral contrast coefficients* and *valleys*, which are related to the shape of the spectrum [24], are also among the results. The first *spectral contrast valley* was present for the Bass and Guitar, and the second *spectral contrast coefficient* for the Vocal.

Dynamic features associated with loudness were present for the Guitar, Vocal and Keys. These are related to the *rms*, *long-term loudness (larm)* [25], *loudness stevens* [26] and *loudness vickers* [27]. For the Bass, *effective duration* [9] was present, which is a global temporal indicator associated to the envelope of an audio segment.

The 33rd *harmonic pitch class profile (HPCP)* was one of

TABLE III
LIST OF INTERPRETATION FEATURES.

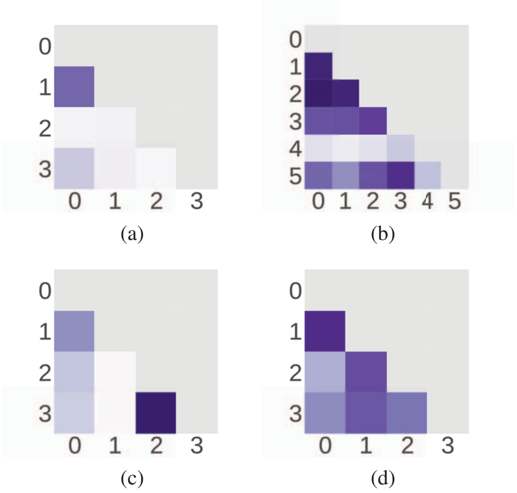
| Group | Name | Pooling |
|---|---|---|
| Bass | 0 - effective duration<br>1 - hpcp (34)<br>2 - spectral energy low<br>3 - barkbands (3)<br>4 - spectral contrast valley (0)<br>5 - barkbands (1) | global<br>variance second derivative<br>mean<br>max<br>max<br>standard deviation |
| Guitar | 0 - spectral energy middle-low<br>1 - spectral energy middle-low<br>2 - spectral energy low<br>3 - rms<br>4 - loudness stevens<br>5 - spectral energy middle-low<br>6 - spectral contrast valley (0) | variance second derivative<br>max<br>mean<br>variance first derivative<br>variance second derivative<br>mean first derivative<br>max |
| Vocal | 0 - spectral contrast coeff. (1)<br>1 - spectral contrast coeff. (1)<br>2 - larm<br>3 - spectral contrast valley (2)<br>4 - larm<br>5 - pitch salience<br>6 - pitch salience | variance<br>standard deviation<br>variance first derivative<br>mean second derivative<br>variance second derivative<br>mean first derivative<br>mean second derivative |
| Keys | 0 - spectral energy<br>1 - larm<br>2 - spectral rms<br>3 - equivalent sound level (leq)<br>4 - loudness stevens<br>5 - rms<br>6 - spectral energy middle-low<br>7 - loudness vickers<br>8 - barkbands (4)<br>9 - spectral energy middle-low<br>10 - barkbands (11)<br>11 - derivative SFX<br>12 - barkbands (5)<br>13 - spectral energy middle-low | max<br>max<br>max<br>max<br>max<br>max<br>max<br>max<br>max<br>standard deviation<br>standard deviation<br>max derivative before max value<br>standard deviation<br>mean first derivative |

TABLE IV
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 |

the selected features for the Bass. The *HPCP* is calculated from the spectral peaks and represents the intensities of various subdivisions of semitone pitch classes [28]. 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 [29].

These spectral, temporal and harmonic features could be an indicator of common practices in stem audio mixing due to the application of audio effects such as EQ, DRC, saturation or pitch correction.



(a) (b)

(c) (d)

Fig. 6. Correlation heatmap among prediction features for (a) Bass, (b) Guitar, (c) Vocal and (d) Keys. Colour intensity represents correlated features.

TABLE V
ACCURACY WITH DIFFERENT CLASSIFIERS AND SET OF FEATURES.

| Inst. | Features | Test score | | |
|---|---|---|---|---|
| | | RF | SVC | LG |
| Bass | 1812 | 0.607 | 0.429 | 0.5 |
| | 6 | 0.679 | 0.75 | 0.423 |
| | 4 | **0.714** | **0.75** | 0.423 |
| Guitar | 1812 | 0.606 | 0.575 | 0.424 |
| | 7 | 0.727 | 0.727 | 0.727 |
| | 6 | **0.788** | **0.758** | **0.758** |
| Vocal | 1812 | 0.666 | 0.5 | 0.5 |
| | 7 | 0.833 | 0.541 | 0.583 |
| | 4 | **0.833** | **0.583** | **0.625** |
| Keys | 1812 | 0.653 | 0.5 | 0.5 |
| | 14 | 0.653 | 0.692 | 0.807 |
| | 4 | **0.692** | 0.653 | 0.730 |

## *B. Correlated features*

From Fig. 6, the Guitar and Keys presented the largest number of correlated variables, whereas the Bass the least and the Vocal only a pair of correlated features. For the Guitar, the highest correlation occurs between variables related to *rms* and *loudness stevens* values. All the features of the Keys seem to be correlated and the maximum correlation is happening between the *middle-low* and the *total spectral energy*. The features for the Bass and Vocal presented a good indicator of uncorrelated variables with the exception of the *pitch salience* features for the Vocal tracks.

The high correlation indicators for Guitar and Keys are associated with the variance between the instruments and their roles within the different genres, i.e. the lead folk electric guitar is processed differently than a backing pop electric guitar. On the other hand, Bass and Vocal indicators of uncorrelation are related to a more uniform processing method between genres. This is also noticeable with the $OOB$ behaviour from Fig. 4 and Fig. 5.

In addition, $f_{int}$ shows a greater number of correlated features. For example, the Keys have $5$ features related to a maximum loudness indicator, while $f_{pred}$ only has one feature related to loudness. This behaviour is shared with the spectral and temporal features of each group of instruments.

### C. Quantitative performance

From Table V it can be seen that the classifiers tended to achieve a better performance with $f_{pred}$ and the highest accuracy was for the Vocal group. The Keys performed best with $f_{int}$ for SVC and LG, although the $f_{pred}$ had a higher test score for RF.

Overall, when discriminating between raw and stem tracks, the RF classifier achieved the highest accuracy with $f_{pred}$. Therefore, we have found the subset of audio features that most consistently describes the mixing of stems from raw recordings. In this way, these features can represent an audio feature space that is being steadily mapped by this process, and then can lead to a prediction model of this transformation.

## VII. Conclusion

In this paper, we determined the sets of audio features that can describe stem audio mixing as a content-based transformation. We have extracted a set of 1812 audio features from Bass, Guitar, Vocal and Keys raw recordings and stems and we have reduced it to 6 or fewer audio features. We compared the performance of different machine learning classifiers when using the entire and the reduced audio feature sets and we showed that the models improved by an average of 12.38%. The features found are related to spectral, dynamic and harmonic audio characteristics which could be associated with EQ, DRC, saturation and pitch-correction audio effects.

In future work, the feature set obtained can be used to train machine learning regression systems that can predict the value of the respective audio characteristics and thus assist the sound engineer during the stem audio mixing process. Also, an additional study could be done to determine the applied combination of audio effects and its relation to the audio features encountered. Similarly, this method can be extended to stereo features in such way that panning procedures are explored. The method can be improved by having a more robust performance to highly correlated features. Finally, an improvement in the selection of raw recordings can also be explored, so that more than one is taken into account during feature extraction.

## References

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

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

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

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

[5] X. Amatriain *et al.*, "Content-based transformations," *Journal of New Music Research*, vol. 32, no. 1, pp. 95–114, 2003.

[6] V. Verfaille, U. Zolzer, and D. Arfib, "Adaptive digital audio effects (a-dafx): A new class of sound transformations," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1817–1831, 2006.

[7] M. A. Martínez Ramírez and J. D. Reiss, "Analysis and prediction of the audio feature space when mixing raw recordings into individual stems," in *143rd Audio Engineering Society Convention*, 2017.

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

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

[10] P. Hamel *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)*, 2011, pp. 729–734.

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

[12] A. Wilson and B. Fazenda, "Variation in multitrack mixes: analysis of low-level audio signal features," *Journal of the Audio Engineering Society*, vol. 64, no. 7/8, pp. 466–473, 2016.

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

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

[15] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[16] ——, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] C. Strobl *et al.*, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[18] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

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

[20] R. M. Bittner *et al.*, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *15th International Society for Music Information Retrieval Conference (ISMIR)*, vol. 14, 2014, pp. 155–160.

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

[22] D. Bogdanov *et al.*, "Essentia: An audio analysis library for music information retrieval." in *14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 493–498.

[23] N. Giordano, "Spectral analysis of musical sounds with emphasis on the piano," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 846–846, 2015.

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

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

[26] R. Teghtsoonian, S. Stevens, and G. Stevens, "Psychophysics: Introduction to its perceptual, neural, and social prospects," *The American Journal of Psychology*, vol. 88, no. 4, p. 677, 1975.

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

[28] E. Gómez, "Tonal description of polyphonic audio for music content processing," *INFORMS Journal on Computing*, vol. 18, no. 3, pp. 294–304.

[29] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *The Journal of the Acoustical Society of America*, vol. 71, no. 3, pp. 679–688, 1982.