ANALYSING SCATTERING-BASED MUSIC CONTENT ANALYSIS SYSTEMS: WHERE'S THE MUSIC?

Francisco Rodríguez-Algarra Centre for Digital Music
Bob L. Sturm Centre for Digital Music
Hugo Maruri-Aguilar School of Mathematical Sciences
Queen Mary University of London, U.K.
{f.rodriguezalgarra, b.sturm, h.maruri-aguilar}@qmul.ac.uk

ABSTRACT

Music content analysis (MCA) systems built using scattering transform features are reported quite successful in the *GTZAN* benchmark music dataset. In this paper, we seek to answer why. We first analyse the feature extraction and classification components of scattering-based MCA systems. This guides us to perform intervention experiments on three factors: train/test partition, classifier and recording spectrum. The partition intervention shows a decrease in the amount of reproduced ground truth by the resulting systems. We then replace the learning algorithm with a binary decision tree, and identify the impact of specific feature dimensions. We finally alter the spectral content related to such dimensions, which reveals that these scattering-based systems exploit acoustic information below 20 Hz to reproduce *GTZAN* ground truth. The source code to reproduce our experiments is available online. [1]

1. INTRODUCTION

Music content analysis (MCA) systems trained and tested in [2] reproduce a large amount of the ground truth of the benchmark music dataset *GTZAN* [18], and are among the "best" reported in the literature [14]. They use support vector machines (SVM) classifiers trained on features extracted from audio by the *scattering transform*, a non-linear spectrotemporal modulation representation using a cascade of wavelet transforms [8]. The mathematical derivation of the scattering transform enforces invariances to local time and frequency shifts, which is a desirable property for music classification tasks. Scattering features are considered to have perceptual relevance [2], and can be related to modulation features [5]. Such features are potentially useful for timbre-related music classification tasks, such as instrument recognition [11], or genre recognition [7].

Reproducing the ground truth of a dataset does not necessarily reflect the ability of a system to address a particular task [12–15]. In this paper, we analyse scattering-based MCA systems to determine why they reproduce so much *GTZAN* ground truth. Our approach involves system analysis and experimental interventions. System analysis involves decomposing an MCA system into its components to understand how each contributes to its overall behaviour. Our system analysis of scattering-based MCA systems in Sec. 2 shows that they use some information from inaudible frequencies, i.e., below 20 Hz [4]. Experimental interventions, on the other hand, involve testing hypotheses about what a system is actually doing by altering some factor to see how system behaviour changes. In Sec. 3, we perform intervention experiments to confirm that scattering-based MCA systems exploit information below 20 Hz to reproduce *GTZAN* ground truth. When we attenuate that information, ground truth reproduction decreases.

We conceive our work here as a case study within the development of an improved systematic methodology for evaluating MCA systems. This is one challenge posed in the Music Information Retrieval (MIR) Roadmap [10], and exemplifies the pipeline in [15]. In Sec. 4 we discuss the implications of our results, and suggest how they might be integrated in a general evaluation framework.

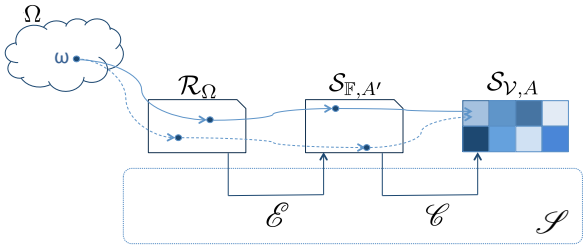


Figure 1. Schematic representation of a music content analysis (MCA) system [16].

2. SYSTEM ANALYSIS

Using the formalism of [16], an MCA system $\mathscr{S}$ maps a recording universe $\mathcal{R}_\Omega$ — a particular realisation of an intangible music universe Ω — to a description universe $\mathcal{S}_{\mathcal{V},A}$. As shown in Fig. 1, this mapping is decomposed into two stages. First, a feature extractor $\mathscr{E}$ maps $\mathcal{R}_\Omega$ to a feature universe $\mathcal{S}_{\mathbb{F},A'}$; then, a classifier $\mathscr{C}$ maps $\mathcal{S}_{\mathbb{F},A'}$ to $\mathcal{S}_{\mathcal{V},A}$.

The environment and definition of the MCA systems in [2] are as follows. $\mathcal{R}_\Omega$ consists of time-domain sig-

[1] https://code.soundsoftware.ac.uk/projects/scatter-analysis

ID	$\mathbb{F}$	Feature Description
a	$\mathbb{R}^{252}$	Mel-frequency spectrogram (84 coefficients, 740-ms frames, 50% overlap), concatenated with first- and second-order time derivatives over the sequence of feature vectors [2]
b	$\mathbb{R}^{85}$	First-order ($l = 1$) time-scattering features (effective sampling rate 2.7 Hz)
c	$\mathbb{R}^{747}$	Second-order ($l = 2$) time scattering features (effective sampling rate 2.7 Hz)
d	$\mathbb{R}^{1574}$	First-order time-frequency scattering features
e	$\mathbb{R}^{1907}$	First-order time-frequency-adaptive scattering features
f	$\mathbb{R}^{2769}$	Third-order ($l = 3$) time scattering features (effective sampling rate 2.7 Hz)

Table 1. Description of $\mathcal{S}_{\mathbb{F},A'}$ (feature universes) used in [2]. A' permits only vector sequences of length 80.

nals of duration about 30 seconds uniformly sampled at $F_s = 22050$ Hz (the sampling rate of *GTZAN*). $\mathcal{S}_{\mathcal{V},A}$ is the set of the 10 *GTZAN* labels. $\mathcal{S}_{\mathbb{F},A'}$ is a space consisting of sequences of 80 elements of a vector vocabulary $\mathbb{F}$. All MCA systems trained in [2] use the same $\mathcal{S}_{\mathcal{V},A}$, the same learning method to build the classifiers, but different $\mathcal{S}_{\mathbb{F},A'}$. More specifically, the semantic rules A' are the same for all systems (sequences of length 80), with only a difference in the feature vocabulary, $\mathbb{F}$. Table 1 describes the six different $\mathcal{S}_{\mathbb{F},A'}$ appearing in [2].

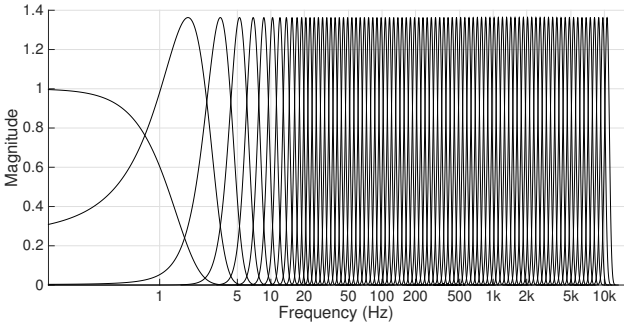
We now analyse the two components of the systems built using first- ("b vectors") and second-layer ("c vectors") time scattering features (see Table 1). Systems built using f vectors can be understood as a further iteration of the process described here. In addition to that, the inclusion of frequency-scattering features (d and e vectors), does not affect our conclusions.

2.1 Feature extractors of b and c ($\mathscr{E}_{\mathtt{b}}$ and $\mathscr{E}_{\mathtt{c}}$)

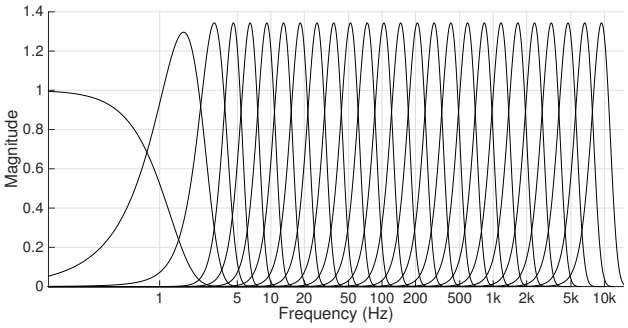
The feature extraction procedure begins by first extending a recording to be of length $2^{21} = 2097152$ samples using what is referred to in the code as "padding" by "symmetric boundary condition with half-sample symmetry": the $N \approx 5 \cdot 2^{17}$ samples of $r \in \mathcal{R}_\Omega$ are concatenated with the same but time-reversed, then concatenated with its first ~ 50000 samples, and its last ~ 50000 samples, and finally the time-reversed samples again. This "padded" signal is then transformed into the frequency domain by the FFT. The complex spectrum is then multiplied by the magnitude response of each of 85 filters of a filterbank designed using a scaling function and dilations of a one-dimensional Gabor mother wavelet with 8 wavelets per octave up to a maximum dilation of $2^{73/8}$. (The bandwidth of the lowest 11 bands are made constant.) Figure 2(a) shows the magnitude responses of the bands of this filterbank (FB1). Each spectrum product is then reshaped — equivalent to decimation in the time-domain —, transformed to the time domain by the inverse FFT, and then windowed to the portion corresponding to r in the padded sequences.

Next, the time-series output of each band of FB1 is rectified, padded using the same padding method as above,

[2] [2] does not actually compute Δ- and Δ-Δ-MFCCs, but instead cyclically time-shifts the sequence of MFCCs ahead and behind by one frame so that the classifier has flexibility in learning a transformation.



(a) Filterbank 1 (FB1)



(b) Filterbank 2 (FB2)

Figure 2. Magnitude responses of the bands in the filterbanks for scattering feature IDs b and c.

transformed into the frequency domain by the FFT, and then multiplied by the magnitude response of each of 25 filters of a filterbank designed with a scaling function and dilations of a one-dimensional Morlet mother wavelet, with 2 wavelets per octave up to a maximum dilation of $2^{23/2}$. Figure 2(b) shows the magnitude responses of the bands of this filterbank (FB2). Each FB2 spectrum product is then reshaped — again, equivalent to decimation in the time-domain —, transformed to the time domain by the inverse FFT, and then windowed corresponding to the original forward-going sequence in the padded sequences (length 80). Finally, $\mathscr{E}_{\mathtt{b}}$ retains only those values related to the DC filter of FB2, and computes the natural log of all values (added with a small positive value). This results in 80 b vectors. For creating 80 c vectors, $\mathscr{E}_{\mathtt{c}}$ takes those FB2 time-series outputs with non-negligible energy, [3] "renormalises" each non-zero frequency band (to account for energy captured in the first layer of scattering coefficients), and takes the natural log of all values (added with a small positive value).

Figure 3 shows the relationship between the dimensions of b and c vectors and the centre frequencies of FB1 and FB2 bands. For display purposes, the bottom-most row is from the scaling function of FB2. The 85 dimensions of a b vector are at bottom, with dimensions [1, 75:85] coming from FB1 bands with centre frequencies below 20 Hz. Dimensions [1, 75:85, 737:747] of a c vector come from such bands. Dimensions [2:12] of a b vector, and [2:12, 86:268] of a c vector, are from FB1 bands with centre frequencies above 4186 Hz (pitch C8).

[3] In fact, not every rectified FB1 band output is filtered by all FB2 bands because filtering by FB1 will remove all frequencies outside its band.

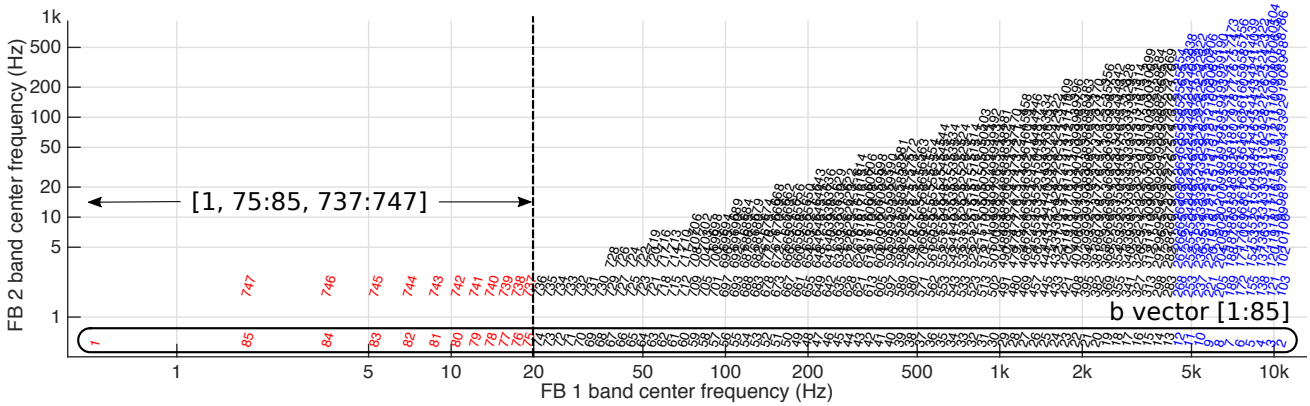


Figure 3. Relationship of b and c vector dimensions to FB1 and FB2 band centre frequencies. Dimensions [1, 75:85] of b vectors, and [1, 75:85, 737:747] of c vectors, are from bands with centre frequencies below 20 Hz.

2.2 Classifier $\mathscr{C}$

Define the number of support vectors of a trained SVM as $|SV|$. Classifiers $\mathscr{C}$ of the MCA systems in [2] are characterised by a set of support vectors $\mathbf{V} \in \mathbb{F}^{|SV|}$, a Gaussian kernel parameter γ, a weight matrix $\mathbf{W} \in \mathbb{R}^{|SV| \times 45}$, and a bias vector $\boldsymbol{\rho} \in \mathbb{R}^{45}$. (45 is the number of pair-wise combinations of the 10 elements in $\mathcal{S}_{\mathcal{V},A}$, i.e., label 1 vs. label 2, label 1 vs. label 3, etc.) $\mathscr{C}$ maps $\mathcal{S}_{\mathbb{F},A'}$ to $\mathcal{S}_{\mathcal{V},A}$ by majority vote from the individual mappings of all elements $f_j \in \mathbb{F}$ of a sequence from $r \in \mathcal{R}_\Omega$ by an SVM classifier $\mathscr{C}'$. $\mathscr{C}'$, thus, maps $\mathbb{F}$ to $\mathcal{S}_{\mathcal{V},A}$ by computing 45 pair-wise decisions by means of $sign(\mathbf{W}^T \mathrm{e}^{-\gamma \mathbf{K}(f)} - \boldsymbol{\rho})$, where $\mathbf{K}(f)$ is a vector of squared Euclidean norm of differences between f and all $v_j \in \mathbf{V}$. $\mathscr{C}'$ then maps f to $\mathcal{S}_{\mathcal{V},A}$ by majority vote from the 45 pair-wise decisions.

The authors of [2] use LibSVM [4] to build $\mathscr{C}'$ using a Gaussian kernel with a subset of the feature vectors (downsampled by 2). They optimise the SVM parameters by grid search and 5-fold cross-validation on a training set. LibSVM uses a 1 vs. 1 strategy to deal with multiclass classification problems, so each support vector receives a weight for each of the nine possible pair-wise decisions involving the class associated with the support vector. The matrix $\mathbf{W}$ contains weights associated with all possible 45 pair-wise decisions. The training of the SVM also generates the vector $\boldsymbol{\rho}$ containing a bias term for each pair-wise decision.

3. SEARCHING FOR THE MUSIC

We now report three intervention experiments we design to answer our question: how are scattering-based MCA systems reproducing so much *GTZAN* ground truth? We adapt the code used for the experiments in [2] (available online [5]). The experiments performed in [2] do not consider the known faults of *GTZAN* [14], so in Sec. 3.1 we reproduce them using two different train-test partitioning conditions. We observe a decrease in performance, but not as dramatic as seen in past re-evaluations [14]. In Sec. 3.2, we replace the classifier $\mathscr{C}$ with a binary decision tree (BDT) trained with different subsets of scattering features. This leads us to identify the impact of specific feature dimensions. The analysis of the feature extractor in Sec. 2.1 allows us to relate such dimensions with spectral bands of the audio signal. In Sec. 3.3, we alter the spectral content of the test recordings and observe how *GTZAN* ground truth reproduction changes. This reveals that these scattering-based MCA systems exploit acoustic information below 20 Hz.

3.1 Partitioning intervention

The benchmark music dataset *GTZAN* contains faults (e.g., repetitions) that can affect the amount of ground truth that an MCA system reproduces [14]. This amount often decreases when we train and test it using a "fault-filtered" partition of *GTZAN*, as done in [6, 14]. This suggests that the faults in the dataset are related to the amount of ground truth reproduced by a system.

While [2] evaluates the performance of the scattering-based MCA systems using 10-fold stratified cross-validation, we employ two different hold-out train-test partitioning conditions. The first is *RANDOM*, which mimics the train-test procedure in [2]: we randomly select 75% of the recordings of each label for the training set, leaving the remaining 25% for the testing set. The second is *FAULT*, which is the "fault-filtered" partitioning procedure in [6], with the training and validation sets merged. This partitioning condition considers various problems of the dataset: we remove 70 replicated or distorted recordings [14]; we then assign by hand 640 recordings to the training set and the remaining 290 to the testing set, avoiding repetition of artists across partitions [9]. Due to memory constraints, we decrease by a factor of 4 the number of scattering features in the pre-computation of the Gaussian kernel of the SVM. This reduces the computational cost without sacrificing much performance. [6]

Table 2 shows the normalised accuracies (mean recalls) of our systems along with those reported in [2] for the six features described in Table 1. We see the differences between the results in [2] and ours in *RANDOM* are small, and most of them within reason considering the standard

[4] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[5] http://www.di.ens.fr/data/software

[6] We acknowledge Joakim Andén and Vincent Lostanlen for their valuable advice.

	Original *GTZAN* recordings			Attenuated [0, 20] Hz	
ID	Reported in [2]	*RANDOM*	*FAULT*	*RANDOM*	*FAULT*
a	82.0 ± 4.2	78.00	53.29	39.20	30.09
b	80.9 ± 4.5	79.20	54.96	31.60	22.42
c	89.3 ± 3.1	88.00	66.46	50.80	44.47
d	90.7 ± 2.4	87.20	68.49	62.40	55.11
e	91.4 ± 2.2	85.60	68.61	64.80	44.52
f	89.4 ± 2.5	83.60	68.32	64.80	53.16

Table 2. Normalised accuracies (mean recall) in *GTZAN* dataset obtained by scattering-based MCA systems in [2] and our systems using *RANDOM* and *FAULT* partitioning conditions, trained and tested with the original *GTZAN* recordings (left) and versions ones with information below 20 Hz (right) attenuated (see Sec. 3.3).

N	*RANDOM*	*FAULT*
1	52.99	51.82
2	65.05	65.27
3	71.42	71.62
4	75.53	75.80
5	79.48	79.74

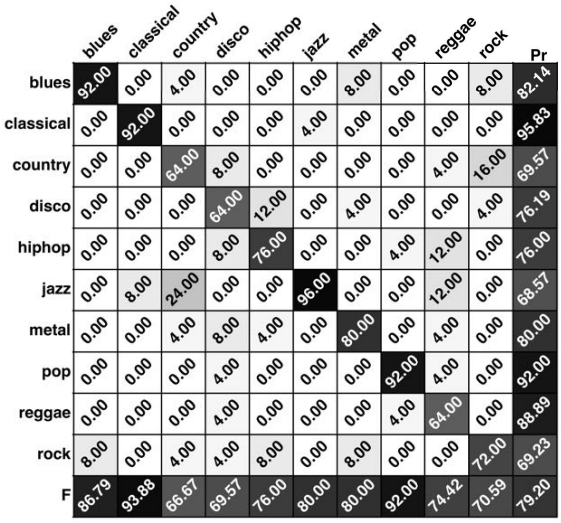
Table 3. Cumulative percentage of variance captured by the N highest principal components of b vectors in the training sets of *RANDOM* and *FAULT* partitioning conditions.

deviations reported in [2]. In *RANDOM*, we see an increase of accuracy when we include second-order scattering features, i.e., b to c. We find that adding depth to the features, however, does not increase further the amount of ground truth reproduced, and even decreases it when we include third-order features (c to f), contrary to what is reported in [2]. Most importantly, we observe a considerable decrease in the amount of ground truth reproduced by all systems between *RANDOM* and *FAULT*. Figures 4(a) and 4(b) show the figure-of-merit (FoM) of the systems trained and tested in *RANDOM* and *FAULT* with b vectors, respectively. We see recalls and F-measures of every label decrease except for "classical", which increase.

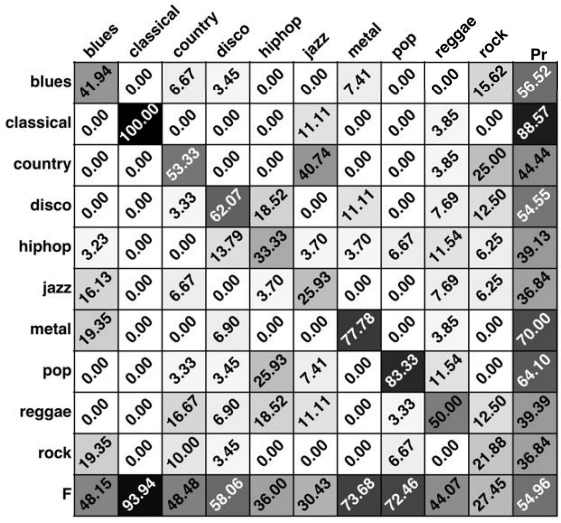
Figure 5 shows the eigenvectors of the first five principal components of first-layer scattering features (b vectors) in the training sets of *RANDOM* and *FAULT*. (Table 3 shows the percentage of variance captured by the first N principal components.) We see large changes in the lowest and highest dimensions of the fourth component. This suggests that these dimensions of the scattering features capture information that differs in both training sets, which may play a role in the performance differences we observe. The characteristics of $\mathscr{C}$ and $\mathscr{C}'$, however, make it difficult to determine the influence that each individual feature dimension (or subset of dimensions) has in the overall performance of a system. For this reason, in Section 3.2 we replace the SVM by a binary decision tree (BDT) classifier, which allows an easier interpretation of $\mathcal{S}_{\mathbb{F},A'}$ and its relationship with $\mathcal{S}_{\mathcal{V},A}$.

3.2 Classifier intervention

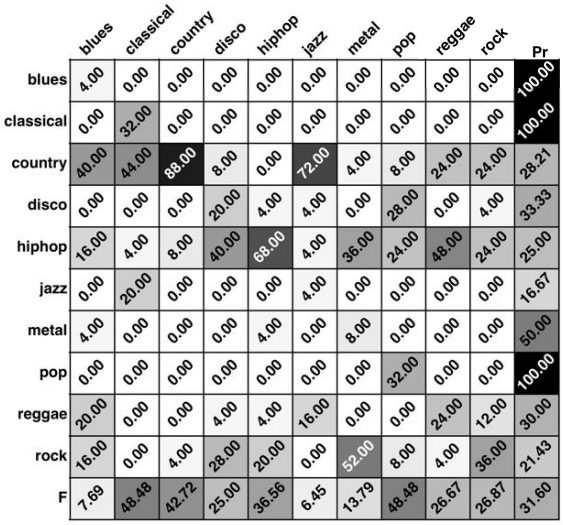
SVM classifiers generate decision boundaries in multi-dimensional spaces. While this can benefit prediction, it hampers their interpretability. In our case, this implies that



(a) *RANDOM*

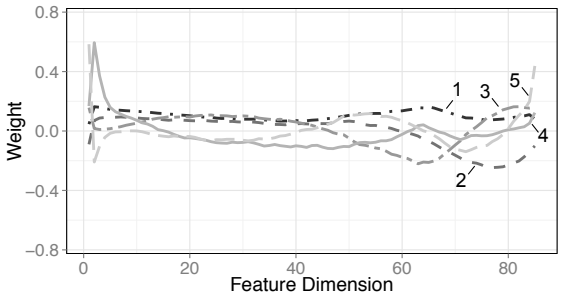


(b) *FAULT*

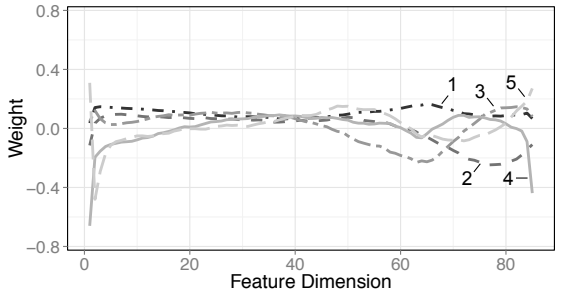


(c) *RANDOM*, information below 20 Hz attenuated

Figure 4. Figure-of-merit (FoM) obtained with b vectors by SVM systems trained and tested in (a) *RANDOM* and (b) *FAULT* (Sec. 3.1), as well as (c) SVM trained in *RANDOM* and tested in recordings with content below 20 Hz attenuated (Sec. 3.3). Column is ground truth, row is prediction. Far-right column is precision, diagonal is recall, bottom row is F-score, lower right-hand corner is normalised accuracy. Off-diagonals are confusions.



(a) *RANDOM*



(b) *FAULT*

Figure 5. Eigenvectors of the first five principal components (labelled) of b vectors in the training sets of (a) *RANDOM* (79.74% of variance captured) and (b) *FAULT* (79.48% of variance captured) partitioning conditions.

ID	*RANDOM*	*FAULT*
a	72.80	45.70
b	71.60	42.35
c	80.00	49.91
d	79.20	46.81
e	79.60	44.77
f	79.20	46.48

Table 4. Normalised accuracies (mean recall) in *GTZAN* for MCA systems built using binary decision tree classifiers using features described in Table 1, trained and tested with *RANDOM* and *FAULT* partitioning conditions.

the relevance of each individual dimension of the scattering feature vectors gets blurred. We now replace the SVM classifiers used in [2] by BDT, consisting of a set of rules defined by linear splits of the feature space one dimension at a time. BDT are considered to be among the easiest learning methods to construct and understand [1], at the cost of potentially less accuracy.

Table 4 shows the normalised accuracies we obtain with MATLAB's BDT classifier, [7] for the two partitioning conditions defined in Sec. 3.1, using the different feature vectors described in Table 1. Clearly, there exists a major difference between the two training conditions, similar to what Table 2 shows for SVM. We see a decrease of around 8 percentage points in the amount of ground truth reproduced by each of the BDT systems in *RANDOM* compared to the SVM systems in Table 2. On the other hand, when training the BDT systems in *FAULT*, we observe falls in performance with respect to *RANDOM* at least as large as those reported in Table 2. This suggests that the amount of

[7] http://uk.mathworks.com/help/stats/classificationtree-class.html

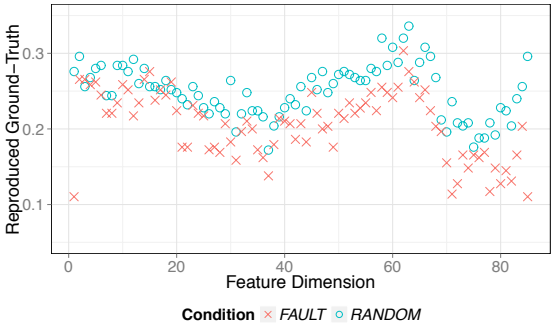


Figure 6. Proportion of ground truth reproduced by BDT classifiers trained with single dimensions of b vectors in *RANDOM* and *FAULT* partitioning conditions.

ground truth reproduced by the systems in both conditions differ due to distinct information being captured by the feature extractors, and not necessarily as an effect of the classification algorithm. The training of the SVM classifier, if anything, appears to mitigate the potential performance decrease in c-f vectors.

We now explore differences in reproduced *GTZAN* ground truth between partitioning conditions by each dimension of the scattering features individually. We thus train and test BDTs with each of the 85 dimensions of b vectors in both *RANDOM* and *FAULT*. Figure 6 shows the classification accuracies we obtain. We see clear differences between conditions, especially in dimensions identified in Sec. 2.1 as belonging to bands close to or outside the limits of normal human hearing (namely [1, 70:85]).

We also explore how much ground truth BDT systems can reproduce using solely information below 20 Hz. BDT systems trained with dimensions 1 and 75:85 of b vectors achieve a 60.40% of normalised accuracy in *RANDOM*, which is close to the performance originally reported in [18] for the *GTZAN* dataset. In *FAULT*, however, the normalised accuracy drops to 22.47%. Adding dimensions 737:747 from c vectors (modulations from FB2 of information below 20 Hz) only marginally increases the performance in both conditions. These results suggest that our scattering-based MCA systems could be exploiting acoustic information from below 20 Hz. We next perform interventions to test this hypothesis.

3.3 Filtering intervention

We now see how the amount of ground truth reproduced by a scattering-based an MCA system changes when we attenuate acoustic information below 20 Hz. We thus apply a fifth-order Butterworth high-pass filter to attenuate all frequencies below 20 Hz by at least 30 dB to the test recordings in both *RANDOM* and *FAULT*. We check that we do not perceive differences between filtered and non-filtered versions. We then use the SVM systems in Sec. 3.1 to predict labels in the filtered test recordings. The two right-most columns of Table 2 show the normalised accuracies we obtain. We clearly see that the figures drop from those reported in Sec. 3.1. In particular, the decrease of accuracy using b vectors in *RANDOM* is close to 50 percentage points, while that using features generated from

deeper scattering layers is smaller but still notable. Systems trained in *FAULT* also suffer in the performance measured.

Figure 4(c) shows the FoM obtained by an SVM trained in *FAULT* with `b` vectors and tested in high-pass filtered recordings. We note that the changes in FoM between Figs. 4(a) and 4(c) do not always match those reported in Sec. 3.1 between Figs. 4(a) and 4(b). More precisely, recall and F-measure decrease instead of increase in "classical", and increase instead of decrease in "country". This suggests that partitioning and information below 20 Hz are distinct factors affecting the amount of ground truth systems reproduce, notwithstanding an interaction between them as suggested by Figs. 5 and 6. Our results allow us to conclude that the scattering-based MCA systems trained and tested in [2] benefit from partitioning and exploit acoustic information below 20 Hz to reproduce a large amount of *GTZAN* ground truth.

4. DISCUSSION

Our analysis in Sec. 2.1 shows how first- and second-layer time-scattering features relate to acoustic information. We see that several dimensions of such features capture information at frequencies below 20 Hz, which is inaudible to humans [4].

We find in the intervention experiments in Secs. 3.1 and 3.2 that partitioning affects the amount of *GTZAN* ground truth scattering-based systems reproduce. Removing the known faults of the dataset and avoiding artist replication across folds leads to a decrease in the FoM we obtain, but to a lesser extent than previous re-evaluations of other MCA systems [6, 14]. We also note that differences between the first principal components of first-layer time-scattering features lay mainly within the dimensions corresponding to frequency bands below 20 Hz.

When we replace the SVM classifier with a BDT (Sec. 3.2), we see differences in the amount of reproduced ground truth similar to those we find for SVM systems between partitioning conditions. This suggests that the distinct acoustic information the scattering features capture causes differences in performance, regardless of the particular learning algorithm employed. Furthermore, we find that BDT systems trained with individual dimensions of first-layer time-scattering features reproduce an amount of *GTZAN* ground truth larger than that expected when selecting randomly. Again, we see differences between partitioning conditions, especially in the dimensions capturing information below 20 Hz. Moreover, we reproduce almost as much *GTZAN* ground truth as the one originally reported in [18] by using a BDT trained in *RANDOM* with only information below 20 Hz. This result suggests that acoustic information below 20 Hz present in *GTZAN* recordings may inflate the performance of MCA systems trained and tested in the benchmark music dataset.

Our system analysis in Sec. 2 and intervention experiments in Sec. 3.1 and 3.2 point toward information present in frequencies below 20 Hz playing an important role in the apparent success of the scattering-based MCA systems we examine. The results of our experiments in Sec. 3.3 clearly reveal that the amount of *GTZAN* ground truth SVM scattering-based systems reproduce decreases when we attenuate that information in test recordings. This implies these systems are using inaudible information. We conclude that the scattering-based MCA systems in [2] exploit acoustic information not controlled by partitioning and below 20 Hz to reproduce a large amount of *GTZAN* ground truth. Machine music listening is an exciting prospect as it complements and even extends human abilities, but we dispute the relevance of acoustic information below 20 Hz to address the problem intended by *GTZAN* [18].

The results of our three intervention experiments suggest a complex relationship between the accuracy measured of a system, the contribution of its feature extraction and machine learning components, and the conditions of the training and testing dataset. We already know that the faults and partitioning of *GTZAN* can have significant effects on an outcome, and that there is an interaction with the components of a system [14, 17]. Our experiments here show for the systems we examine that acoustic information below 20 Hz can greatly affect an outcome, and that this interacts with the components of a system and the dataset partitioning. This thus calls into question the interpretation of the results reported in [2] (column 2 of Table 2) as unbiased estimates of system success. In future work, we will specify more complex measurement models, e.g., [17].

Understanding how and why a system works is essential to determine its suitability for a specific task, not to mention its improvement. Our work here demonstrates the use of system analysis and the intervention experiment to address this problem. For instance, our conclusions suggest modifying the FB1 filterbank in the scattering features extractor to avoid capturing information below 20 Hz. They also suggest removing information below 20 Hz from any element of $\mathcal{R}_\Omega$ as a pre-processing step before training an MCA system, if relevant.

5. CONCLUSION

In this paper, we report several steps we followed to determine what the scattering-based MCA systems reported in [2] have actually learned to do in order to reproduce the ground truth of *GTZAN*. We show how performing system analysis guides our design of appropriate intervention experiments. The results lead us to conclude that these MCA systems benefit not only from the partitioning of the dataset, but also from acoustic information below 20 Hz.

Our work here constitutes steps toward a holistic analysis of MCA systems — an action point for MIR evaluation identified in [10]. Our ultimate goal is to help develop a general MIR research pipeline that integrates system analysis and interventions, and is grounded in formal principles of statistical design of experiments, e.g., [3]. Such a pipeline will provide a solid empirical foundation upon which to build machine music listening systems and technologies [15].

6. REFERENCES

[1] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, Cambridge, MA, USA, 3rd edition, 2014.

[2] J. Andén and S. Mallat. Deep Scattering Spectrum. *IEEE Transactions on Signal Processing*, 62(16):4114–4128, 2014.

[3] R. A. Bailey. *Design of Comparative Experiments*. Cambridge University Press, 2008.

[4] A. Chaudhuri. *Fundamentals of Sensory Perception*. Oxford University Press, 2011.

[5] T. Chi, P. Ru, and S. A. Shamma. Multiresolution Spectrotemporal Analysis of Complex Sounds. *Journal of the Acoustical Society of America*, 118(2):887–906, 2005.

[6] C. Kereliuk, B. L. Sturm, and J. Larsen. Deep Learning and Music Adversaries. *IEEE Transactions on Multimedia*, 17(11):2059–2071, 2015.

[7] C. Lee, J. Shih, K. Yu, and H. Lin. Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features. *IEEE Transactions on Multimedia Multimedia*, 11(4):670–682, June 2009.

[8] S. Mallat. Group Invariant Scattering. *Communications on Pure and Applied Mathematics*, LXV:1331–1398, 2012.

[9] E. Pampalk, A. Flexer, and G. Widmer. Improvements of Audio-Based Similarity and Genre Classification. In *Proc. 6th International Society for Music Information Retrieval Conference (ISMIR'05)*, pages 628–633, London, UK, September 2005.

[10] X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez, F. Gouyon, P. Herrera, S. Jordá, O. Paytuvi, G. Peeters, H. V. Schluter, and G. Widmer. *Roadmap for Music Information Research*. The MIReS Consortium, 2013.

[11] K. Siedenburg, I. Fujinaga, and S. McAdams. A Comparison of Approaches to Timbre Descriptors in Music Information Retrieval and Music Psychology. *Journal of New Music Research*, 45(1):27–41, January 2016.

[12] B. L. Sturm. Classification Accuracy Is Not Enough. *Journal of Intelligent Information Systems*, 41(3):371–406, 2013.

[13] B. L. Sturm. A Simple Method to Determine if a Music Information Retrieval System Is a "Horse". *IEEE Transactions on Multimedia*, 16(6):1636–1644, 2014.

[14] B. L. Sturm. The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval. *Journal of New Music Research*, 43(2):147–172, 2014.

[15] B. L. Sturm. Revisiting Priorities: Improving MIR Evaluation Practices. In *Proc. 17th International Society for Music Information Retrieval Conference (ISMIR'16)*, New York, NY, USA, August 2016.

[16] B. L. Sturm, R. Bardeli, T. Langlois, and V. Emiya. Formalizing the Problem of Music Description. In *Proc. 15th International Society for Music Information Retrieval Conference (ISMIR'14)*, pages 89–94, Taipei, Taiwan, October 2014.

[17] B. L. Sturm, H. Maruri-Aguilar, B. Parker, and H. Grossman. The Scientific Evaluation of Music Content Analysis Systems: Valid Empirical Foundations for Future Real-World Impact. In *Proc. ICML Machine Learning for Music Discovery Workshop*, Lille, France, July 2015.

[18] G. Tzanetakis and P. Cook. Musical Genre Classification of Audio Signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–301, 2002.