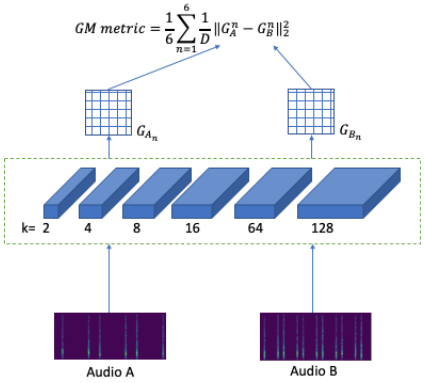

Figure 1. Gram matrix param-sensing metric computation block diagram consisting of an ensemble of six 1 layer 1d-CNNs, with different kernel sizes, k. Each CNN has 512 filters, and the weights are randomly initialized.

architecture used by Antognini et al. [5] for the Gram matrix computation, as shown in Figure 1. It consists of an ensemble of 6 (N=6) single-layered CNNs. We used spectrograms of the audio texture clips as the input to the network computed with 512 FFT bins and hop-size of 128 samples. Unlike images, we consider the spectrograms as one-dimensional input features with the frequency bins as the number of channels. We therefore use a one-dimensional convolution, where each CNN has a convolutional kernel k_n with a different width, in particular 2, 4, 8, 16, 64, 128. The different kernel sizes capture statistics over multiple time scales. We used $F = 512$ filters randomly initialized from a normal distribution (with no training) thus producing six 512×512 dimensional Gram matrices. Details are available in Supplementary Material [1] .

3.2.2 Gram matrix Cos Metric (GMcos)

We compute the cosine distance between the Gram matrices (instead of mean square error) as,

$$GMcos = 1 - \frac{G_A^n \cdot {G_B^n}^T}{||G_A^n||_2 ||G_B^n||_2} \tag{3}$$

where, G_A^n and G_B^n are flattened Gram matrices of audio clips A and B respectively for the n^{th} CNN in an ensemble of N CNNs.

3.2.3 Accumulated-Gram Metric (AGM)

Gram matrices for audio textures are usually sparse (see Supplementary Material (Section 1.) for more info). Neto et al. [18] used a compact version of the Gram matrices to predict the class label of a given sample. Similarly, for a metric that captures the essence of the Gram matrices generated from an audio texture clip, we compute a *Gram vector* corresponding to the clip by accumulating the values from its six Gram matrices, and then aggregating the rows over all the Gram matrices to get a compact one dimensional summarizing vector of length 128. For more details, please refer to Supplementary Material.

[1] https://animatedsound.com/ismir2022/metrics/supplementary/main.html

We define Accumulated-Gram metric (AGM) as the dot product of the difference of the two Gram vectors g_A and g_B corresponding to the two audio clips A and B as

$$AGM = (g_A - g_B) \cdot (g_A - g_B)^T \tag{4}$$

3.3 Cochlear Param-Metric (CPM)

McDermott and Simoncelli [1] use a 3-step texture model to generate a set of statistics of an audio texture as synthesis parameters. A filterbank with a set of cochlear filters is first applied on the incoming sound to decompose the sound to many sub-bands. Then, sub-band envelopes are computed and compressed. And finally a filterbank with a set of modulation filters is applied on each compressed subband envelope. They use seven sets of time-average statistics of nonlinear functions of either the envelopes or the modulation bands as a representation of textures, which includes marginal statistics, variance, and correlations. During synthesis, a white noise signal is iteratively modified to minimize the distance between the synthesised and target sound statistics. Our experiments are based on McWalter's implementation [4] of the [1] algorithm. An overview of the method, implementation details, and statistical parameters are summarized in Supplementary Material (Section 2.).

We use the statistical parameters calculated by this algorithm as a representation for the Cochlear Param-Metric (CPM). For each sound x, we calculate the seven sets of statistics S^i where $i = [1...7]$, and compute CPM as

$$CPM = \sum_{i=1}^{7} D(S_A^i, S_B^i) \tag{5}$$

where S_A^i is a flattened vector of the i^{th} statistics set of sound clip A, and $D(x1, x2)$ calculates the cosine distance between two vectors.

4. EXPERIMENTAL SETUP

4.1 Dataset

We use a subset of audio textures from [9] as summarized in Table 1. There are 13 texture types, each of which has a collection of examples that are determined by a set of variable control parameters. For example, various *windchimes* sounds are determined by parameters *chimesize* and *strength*. For our experiments only one control parameter varies at a time. For example, for the set of audio textures *windchimes-strength*, we generate 11 audio files with a varying strength parameter, with the first file having the lowest strength parameter value and the 11th file having the highest strength value. We generate 10 files for each these 11 setting, where all the 10 versions have the same parameter settings but are different instances. All the audio files being used are between 1.5 and 2 seconds long.

The *pitched* sound group consists of the frequency modulation, windchimes, chimes (without the wind sound), and feedback noise (FB noise). FB noise has the parameter 'pitchedness' that changes the texture from a noisy

Table 1. Dataset overview

Group	Texture Type	Parameters
Pitched	FM	modulation frequency (FM-mf) carrier frequency (FM-cf) modulation index (FM-mi)
	Windchimes	chimesize (windchimes-size) strength (windchimes-strength)
	Chimes	chimesize (chimes-size) strength (chimes-strength)
	FB Noise	pitchedness (fbnoise-pitchedness)
	Nsynth brass	pitch (nsynth-pitch)
Rhythmic	Pops	rate (pops-rate) center freq (pops-cf) irregularity (pops-irreg)
	Chirps	rate (chirps-rate) center freq (chirps-cf) irregularity (chirps-irreg)
	Tapping	rate (tapping-rate) relative phase (tapping-relphase)
	Drum break	tempo (drum-tempo) reverb (drum-rev)
Others	Wind	gustiness (wind-gust) howliness (wind-howl) strength (wind-strength)
	Waterfill	fill-level (water-fill)
	Bees	center frequency (bees-cf) busy-body (bees-busy)
	Applause	rate (applause-rate) no. of clappers (applause-clappers)

signal to a pitched sound. Under the *rhythmic* group are pops, chirps, tapping, and drum-break. The tapping sound is similar to the "tapping 1-2" sound in the collection of sounds in [1], and the drum break sound set is recorded and then processed with varying tempo and reverb parameters. The *others* group includes wind, water filling a container, buzzing bees, and applause. Most sounds are synthesized [2], except the waterfill, Nsynth brass, and drum-break textures, which are recorded or are manipulated versions of a real recording.

4.2 Subjective Evaluation

Listening tests were conducted to quantify listeners' sensitivity to control parameter changes based on their ability to identify the order and proximity of audio samples w.r.t two selected reference samples as well as each other. This evaluation was done to provide a perceptual baseline for our objective metrics comparison. Each audio trial consisted of 7 audio samples - 2 references and 5 test samples. The 2 references were selected for each texture with the control parameters (as in the Parameters column in table 1) set to 0 (left endpoint) and 1 (right endpoint). The 5 test samples are from the same texture selected with parameters between 0 and 1. A total of 9 such texture-trials were created, each with a different combination of control parameter settings for the 5 test audio samples. This was done to ensure that all the 9 control parameters values between 0 and 1 are included and are uniformly distributed across trials. The references, test samples, and the sequence of the individual texture-trials were randomized while conducting the test.

An interface was developed specifically for this test with the goal to intuitively convey the task details to the listeners. First they listened to the two references and then to the individual samples in the test. Then they were asked to create an arrangement by positioning thumbs on a slider corresponding to each test sample depending on how they perceived the order and distance of each sample w.r.t. the two references. Adjustments could be made by listening to their arrangement until they were satisfied. The listening test interface can be viewed on our webpage [3]. Amazon Mechanical Turk (AMT) was used to collect 30 responses per trial from a total of 348 unique participants.

[2] Dataset Appendix: https://animatedsound.com/ismir2022/metrics/appendix_dataset/index.html

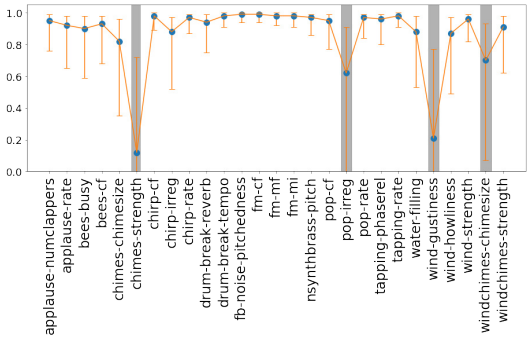

Figure 2. Correlation of perceptual distances with control parameters used to generate the textures. Grey bars indicate textures with no significant correlation.

5. EXPERIMENTS AND RESULTS

In this study, we experimentally address two main questions: 1) Are the objective metrics consistent for texture instances generated with the same parameters (Section 5.2)? 2) Are the objective metrics sensitive to parameter variations (Section 5.3)? We evaluate each metric by comparing them with the subjective responses.

5.1 Subjective tests

Figure 2 shows the correlation coefficients for the subjective perceptual distance captured w.r.t the synthesis control parameters. Parametric variations for chimes-strength, pop-irreg, wind-gustiness and windchimes-chimesize did not result in significant correlations primarily due to wide variance in human ratings, and are excluded from further comparison with objective metrics. Based on the distance measures, we derive rank ordering for the audio samples along the parametric dimension to compare with metrics. See the Supplementary Material for summary rank-order plots obtained from the listening tests for all textures under consideration in this paper. A selection of sounds used in our listening tests can be auditioned on our webpage.

5.2 Metric Consistency

Metric consistency means that the distance between two texture instances with the same parameter settings is small. We analyse consistency in terms of relative mean for three texture types - FM, pops, and water filling. The metric values are computed over one hundred comparisons between two parametrically identical textures. Relative mean is the single parameter mean with respect to the maximum average mean value of comparisons across different parameter settings for the texture as computed in Section 5.3.

[3] https://animatedsound.com/ismir2022/metrics/

Table 2. Metric Consistency in terms of relative mean in % for three texture types - FM, pops, and water filling. Lower is better. The best two in every row are highlighted in bold.

Text-Param	L2	FAD	CPM	GM	GMcos	AGM
FM-cf	54.04	5.69	8.05	**0.07**	**0.02**	0.12
pops-rate	21.36	5.27	**3.44**	6.81	**2.46**	4.28
water-fill	61.09	40.60	23.09	16.19	**7.83**	**12.17**

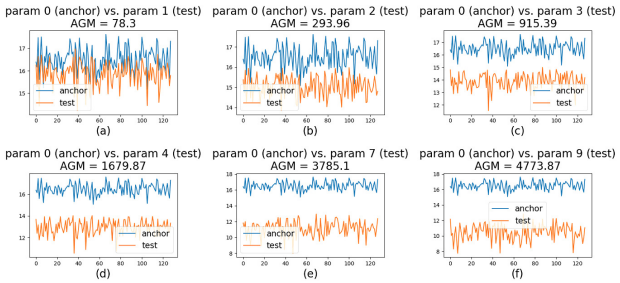

Figure 3. Gram vectors and AGM values of anchor for applause sounds with an increasing number of clappers from (a) to (f) compared to the anchor sound of a single clapper.

The results are shown in Table 2. The L2 metric shows a high relative mean value for all three texture types which confirms the understanding that a spectrogram-based distance measurements does not capture the statistics that define a texture. It is too sensitive to the "irrelevant" variations between similarly parameterized textures.

Although the relative mean of FAD is not as high as that of L2, it is higher than the Gram matrix based metrics, because the embeddings extracted from VGGish provide a holistic representation of the audio including both the overall statistics and temporal structure. The effect of the exact temporal structure of the sound on the metric is reduced but still present. The statistical metrics (CPM and GM-based) exhibit low relative mean values (good consistency), especially GMcos. The waterfill texture shows higher relative mean values across all metrics. This is at least in part because for real recordings, even for short durations, the fill-level is never constant while filling a container.

5.3 Metric Sensitivity to Parameter Variation

To further investigate the sensitivity of the objective metrics to parameter variations, we fix one texture example to a low parameter value (0), called the *anchor*, and compare it with nine *test* clips of the same texture-type but increasing parameter values (1-9). We compute this over 10 versions of the same anchor and test parameter settings, and average over them for the metrics L2, CPM, GM, GMcos, and AGM. However, to calculate FAD, we compute Fréchet distance between normal distributions of embeddings of the 10 versions, instead of between the embeddings of individual audio files (Eq. 1). An example of the sensitivity of the 1×128 dimensional AGM Gram vector to parameter variations is shown in Figure 3. The Gram vectors of a reference (anchor) applause audio texture clip that has the number of clappers parameter set to 1, is compared with six test audio clips of applause textures with increasing number of clappers. The divergence between the Gram vectors of the anchor and test sounds grows as the number of clappers is increased.

Table 3. Pearson's Correlation between the avg human rank orders and the distance from the anchor computed by the objective metrics. The best two values in every row are in bold. Correlation values >=0.5 are green, <0.5 orange.

Texture-Param	L2	FAD	CPM	GM	GMcos	AGM
Pitched						
FM-mf	**0.99**	0.94	**0.99**	**1.00**	**0.99**	0.64
FM-cf	**0.99**	0.71	**0.99**	**1.00**	**0.99**	0.97
FM-mi	**0.99**	0.75	0.94	**0.99**	**1.00**	**0.99**
windchimes-strength	**0.97**	**0.97**	0.82	**0.95**	**0.95**	0.62
chimes-size	0.28	**0.88**	**0.92**	0.20	0.20	0.07
fbnoise-pitchedness	**1.00**	0.75	0.98	**1.00**	**0.99**	**1.00**
nsynth-pitch	0.09	-0.45	**0.73**	**0.23**	0.09	-0.50
Rhythmic						
pops-rate	**0.98**	0.85	0.80	0.79	0.89	**0.94**
pops-cf	**0.98**	0.96	**0.99**	**0.99**	**0.99**	**0.98**
chirps-rate	**0.97**	**0.94**	**0.94**	0.84	0.88	0.89
chirps-cf	**0.99**	**0.98**	**0.98**	**0.98**	**0.98**	0.97
chirps-irreg	0.81	**0.95**	**0.94**	0.90	0.92	0.90
tapping-rate	**1.00**	0.90	0.91	0.64	**0.94**	0.76
tapping-relphase	0.88	**0.96**	**0.98**	0.94	0.95	0.76
drum-tempo	0.08	**0.82**	**0.97**	**0.97**	0.63	0.81
drum-rev	**0.93**	0.85	0.81	0.90	0.81	**0.94**
Others						
wind-howl	**0.92**	**0.92**	**0.92**	0.80	**0.92**	0.86
wind-strength	**0.96**	0.86	**0.88**	0.87	**0.88**	0.56
water-fill	0.39	0.32	0.41	**0.75**	**0.73**	-0.27
bees-cf	**0.97**	0.90	**0.98**	**0.97**	**0.98**	0.80
bees-busy	-0.21	**0.89**	**0.89**	-0.34	-0.43	**0.92**
applause-rate	0.66	0.83	**0.86**	0.66	0.78	**0.90**
applause-clappers	**0.97**	0.94	0.93	**0.99**	**0.99**	**0.99**

Figure 4 shows the plots of all the metric values for three textures compared with human responses in terms of average rank order. The Pearson's correlation between the objective metrics and the subjective responses are presented for all textures in Table 3, and their corresponding plots are available in the Supplementary Material.

5.3.1 Metric Performances

The L2 metric is again a poor performer. This is particularly clear for drumbreak-tempo textures as can be observed in the L2 column in Figure 4(b). FAD performs better than Gram matrix based measures for chimes-size, but shows worse performance for FM-cf, FM-mi, FBnoise-pitchedness, and waterfill (Table 3). FAD performs well for most of the rhythmic textures except pop-rate, drum-tempo, and drum-rev. Since FAD preserves some information about exact spectro-temporal structure, the result is variable sensitivity to parameters (Figure 4, FAD row).

The statistics designed by [1] are known to synthesize good quality sounds for natural textures such as bees and wind which is clearly reflected in our results. However, these statistics are also known for low realism scores for synthesized pitched sounds such as windchimes, and rhythmic sounds such as drum break. Compared to the Gram matrix based metrics, CPM performs well except for windchimes-strength, pop-rate, drum-rev, and water-fill.

When an event rate varies, the features captured by the Gram matrix may reflect some individual events, especially for shorter kernel sizes. In such cases, AGM captures the summary of those statistics instead of individual events, thereby showing a higher parameter sensitivity than metrics based on the entire matrix. This trend can be seen

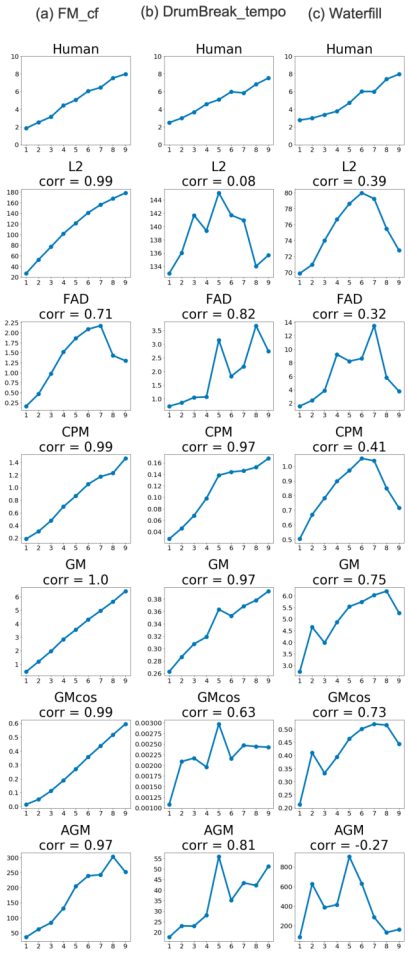

Figure 4. Trends of human responses and objective metrics to parameter variations for three texture-param combinations, (a) FM with carrier frequency variable parameter, (b) Drum-break with tempo a variable parameter, and (c) Waterfill with fill-level parameter, along with the Pearson's correlation between the objective metric values and the subjective responses.

in applause-rate, pop-rate, tapping-rate, chirp-rate, as well as bees-busy (increase in busy body parameter sounds like increase in the rate of buzzing vibrations), where AGM outperforms GM.

The cases where GM outperforms AGM are FM-mf, bees-cf, windchimes-strength, waterfill, and wind-strength, where the parameter variation occurs in the frequency domain, i.e. the center frequency in bees-cf, wind and windchimes-strength, modulation frequency in FM-mf, and resonant frequencies in water-fill. AGM summarization loses information about the statistics in the frequency domain, and this suggests a future study where methods such as eigenvalue decomposition might be better for extracting key information from the Gram matrix.

5.3.2 Some curious cases

Water-fill shows an interesting trend in most of the metrics, where there's a gradual increase, and then a decrease in the metric values (Figure 4(c)), possibly because the resonances of the container wrt changing water and air column heights cause the audio textures of an almost full container to have some statistical similarities with that of an almost empty container. That is, the high-level fill-level parameter has a non-linear affect on the resonant frequencies of the system. For a sound such as water-fill where multiple perceptual dimensions are varying, GM and GMcos show higher correlation with humans than others.

Nsynth-pitch is a curious example where none of the metrics perform well except CPM. The sustained portion of musical tones are atypical as textures because of the lack of temporal variation. Each spectrogram frame is almost exactly the same. Also, as the pitch of the brass instrument increases, the harmonics also change. The Gram matrix summarizes the temporal statistics through the 1D audio representations, but not the frequency statistics, thereby showing a noisy performance.

6. DISCUSSION

We explore a variety of metrics for use with audio textures for their sensitivity to controlled parametric variation. However, in a realistic situation, there may effectively be multiple simultaneous dynamic parameters. One such example in our dataset was the waterfill texture-type, where the fill parameter maps to both rising and falling resonant frequencies. We saw that most of the metrics were responding to a combination of fill parameter and resonances. Further investigation is required to understand how metrics behave in such complex realistic situations.

The systematic parameter variation we used creates textures perceptually close to each other. Further study is required to understand the behavior of these metrics for cross texture-type comparisons, for example, bees compared to water. Such a study would help in mapping audio textures to a perceptual space, the way musical instrument space has been mapped in previous work [19, 20].

Our perceptual study involved placing textures within a 1D space between two reference textures, but there is an inherent lack of an absolute perceptual frame of reference for the meaning of control parameter variation. A systematic study is needed to understand perceptual "just noticeable differences" in parameter variations for complex sounds. Different parametric variations could then be compared on a unified scale.

7. CONCLUSIONS

We present a comprehensive study on the parameter change sensing property of various existing audio evaluation metrics as well as three potential audio statistical and deep-feature based metrics [4]. CPM and FAD emerge as the best metrics, while GM based metrics show promising results. This shows the potential of these deep-features for the purpose of evaluation of audio textures. This study is a fruitful first step towards understanding audio textures, metric design for audio textures, building better synthesis models of this rich and complex class of sounds, and generally toward mapping the space of audio textures.

[4] Code base: `https://github.com/chitralekha18/ParamSensitiveMetrics`

8. REFERENCES

[1] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[2] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1068–1077.

[3] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms." in *INTERSPEECH*, 2019, pp. 2350–2354.

[4] R. McWalter and J. H. McDermott, "Adaptive and selective time averaging of auditory scenes," *Current Biology*, vol. 28, no. 9, pp. 1405–1418, 2018.

[5] J. M. Antognini, M. Hoffman, and R. J. Weiss, "Audio texture synthesis with random neural networks: Improving diversity and quality," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3587–3591.

[6] C. S. Sastry and S. Oore, "Detecting out-of-distribution examples with in-distribution examples and gram matrices," *NeurIPS Workshop on Safety and Robustness in Decision Making*, 2019.

[7] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[8] D. Ulyanov and V. Lebedev, "Audio texture synthesis and style transfer," *[Blog post]. Available from: https://dmitryulyanov. github. io/audio-texture-synthesis-and-style-transfer/[accessed 16 Jan 2022]*, 2016.

[9] L. Wyse and P. T. Ravikumar, "Syntex: parametric audio texture datasets for conditional training of instrumental interfaces." *International Conference on New Interfaces for Musical Expression*, 4 2022, https://nime.pubpub.org/pub/0nl57935. [Online]. Available: https://nime.pubpub.org/pub/0nl57935

[10] J. Bruna and S. Mallat, "Audio texture synthesis with scattering moments," *arXiv preprint arXiv:1311.0407*, 2013.

[11] J. Andén, V. Lostanlen, and S. Mallat, "Joint time–frequency scattering," *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3704–3718, 2019.

[12] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Advances in neural information processing systems*, vol. 28, pp. 262–270, 2015.

[13] H. Caracalla and A. Roebel, "Sound texture synthesis using ri spectrograms," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 416–420.

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

[15] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," *arXiv preprint arXiv:1902.08710*, 2019.

[16] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of multivariate analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[17] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[18] A. J. Neto, A. G. Pacheco, and D. C. Luvizon, "Improving deep learning sound events classifiers using gram matrix feature-wise correlations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3780–3784.

[19] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *the Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, 1977.

[20] P. Esling, A. Chemla-Romeu-Santos, and A. Bitton, "Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces." in *ISMIR*, 2018, pp. 175–181.

STABILITY OF SYMBOLIC FEATURE GROUP IMPORTANCE IN THE CONTEXT OF MULTI-MODAL MUSIC CLASSIFICATION

Igor Vatolkin
TU Dortmund University
Department of Computer Science
igor.vatolkin@udo.edu

Cory McKay
Marianopolis College
Department of Liberal and Creative Arts
cory.mckay@mail.mcgill.ca

ABSTRACT

Multi-modal music classification creates supervised models trained on features from different sources (modalities): the audio signal, the score, lyrics, album covers, expert tags, etc. A concept of "multi-group feature importance" not only helps to measure the individual relevance of features of a feature type under investigation (such as the instruments present in a piece), but also serves to quantify the potential for further improving classification by adding features from other feature types or extracted from different kinds of sources, based on a multi-objective analysis of feature sets after evolutionary feature selection. In this study, we investigate the stability of feature group importance when different classification methods and different measures of classification quality are applied. Since musical scores are particularly helpful in deriving semantically meaningful, robust genre characteristics, we focus on the feature groups analyzed by the jSymbolic feature extraction software, which describe properties associated with instrumentation, basic pitch statistics, melody, chords, tempo, and other rhythmic aspects. These symbolic features are analyzed in the context of musical information drawn from five other modalities, and experiments are conducted involving two datasets, one small and one large. The results show that, although some feature groups can remain similarly important compared to others, differences can also be evident in various applications, and can depend on the particular classifier and evaluation measure being used. Insights drawn from this type of analysis can potentially be helpful in effectively matching specific features or feature groups to particular classifiers and evaluation measures in future feature-based MIR research.

1. INTRODUCTION AND RELATED WORK

There are music research scenarios where manually designed features can have value, and where gaining understanding of which features are particularly effective for various classification tasks can be of central importance. Researchers seeking to gain domain knowledge can be interested in more than just optimizing performance. Those working on composer attribution or genre classification, for example, might be interested in learning about the specific qualities that delineate musical style, so comparing the performance of different musically meaningful features can provide important insight. There can also be situations where only very limited training data are available, such as when there is only so much extant music by a given composer, and where data augmentation techniques can only improve matters so much. In such situations, deep learning that self-learns features can be less useful, and handmade features become valuable, as do approaches for selecting more promising features when many are available.

Multi-modal features drawn from a variety of data sources can be of particular interest. Although there can be redundancy across types of musical data (e.g., both audio and symbolic data specify pitch), such information is not always equally accessible (e.g., pitch is harder to extract from dense polyphonic audio), and there are other times when different source types contain complementary information (e.g., album art can provide cultural information that is inaccessible from audio or scores). Multi-modal research often plays an essential role in musicology, and involving sources as diverse as concert programs, manuscript illuminations, critical accounts, contemporary visual portrayals, and scores themselves. MIR research can and has similarly taken advantage of multi-modal information, both for improving performance in an engineering sense and in learning more about the connections between different types of musical data in a scientific or musicological sense. There is a rich MIR literature involving two modalities, but relatively few studies have considered more [1–6], and relatively few multi-modal public datasets are available (standouts are described in [5, 7–10]). There are also several papers that provide useful summaries of multi-modal MIR research [11–14].

Of course, with more types of data come more features, and the curse of dimensionality becomes a concern. Although handmade features can help, since they tend to represent information more concisely than raw data, too many can still present problems when training data are limited. Although there is substantial literature on dimensionality reduction (e.g., [15]), and important related MIR studies [4, 16–23], some approaches sacrifice feature independence (e.g., PCA), which compromises interpretability,

and feature selection approaches can be sensitive to particular classifier and evaluation methodologies.

So, it can be of value to get a sense of how consistently features perform across different classification and evaluation methodologies, in order to gain a deeper musical understanding of which features and groups of features are most important in delineating classes of interest. This paper is concerned with exactly this: attempting to measure the stability in importance of various features for a given music classification task, which can ultimately help to avoid overfitting the choice of features (or modalities), and can provide insights into underlying musical meaning.

We focus here on symbolic features, as they tend to be particularly interpretable, especially when collaborating with musicologists or theorists. This research maintains a fundamentally multi-modal character, as it also uses information from five other source types to ground feature stability measurements in a broader context: we used the same audio, album cover image, playlist co-occurrence, semantic tag, and lyric text feature data as [10]. These data are particularly useful because they include both a small but clean dataset and a large but noisy dataset, permitting stability to be measured in both scenarios. All symbolic features were extracted with jSymbolic [24], which provides access to many features usefully grouped into feature types whose relative stability can be compared.

We selected genre classification as the test domain for this research, but there is nothing about the techniques we propose that is specific to genre. Furthermore, the same techniques could just as easily be refocused on modalities other than symbolic music. It is also important to note that there are fundamental concerns related to the evaluation of musical classification [25–28]; although there is insufficient space to detail these here, this is essential reading for MIR researchers working in classification.

2. MULTI-GROUP FEATURE IMPORTANCE

We have proposed the concept of *multi-group feature importance* in [10], as an extension of earlier work [29–31]. A feature selection scenario based on two objectives was constructed, with the goal of simultaneously minimizing balanced relative error m_{BRE} (defined as a mean of relative errors for positive and negative instances) and maximizing the share of features g_i belonging to a given group i under investigation, such as rhythmic descriptors. This approach makes it possible to answer not only the questions "what is the lowest error rate achieved by rhythmic features in predicting a musical category?" and "which are the best rhythmic features for the current prediction task?", but also "how can performance be improved when features of other groups, domains, or modalities are also introduced?"

While the formal details of "importance" and the mathematical backing of multi-objective optimization are left to [10], we will briefly illustrate the core ideas here with the help of Figure 1. The connected circles in Subfigure (a) show a *non-dominated front* after feature selection, which simultaneously maximizes the share of rhythmic features g_{RHYTHM} and minimizes m_{BRE}, when features from all six modalities are considered (i.e. not just other symbolic features). Each circle corresponds to a feature set that is not *dominated* by any other feature set. This means that no other feature sets have been identified after feature selection that are "better" than the one under consideration, in combined terms of the two measures being optimized (i.e., using *a greater share* of rhythmic descriptors and achieving *smaller* classification error at the same time). For instance, the feature set in the upper right corner contains only rhythmic descriptors ($g_{RHYTHM} = 1.0$), but $m_{BRE} = 0.4236$ is rather high. No other feature sets that contain only rhythmic descriptors have lower errors, meaning that they are dominated by the upper right corner set, and are not shown in the plot. The feature set in the bottom left corner has the smallest $m_{BRE} = 0.0139$, but has only 37.86% rhythmic descriptors. Other circled feature sets between these corner solutions consist of various trade-offs between the two measures, and are also not dominated by any other found feature sets. The non-dominated fronts in the figures are constructed after ten feature selection repetitions and an independent evaluation on the reserved test set created with music tracks used neither for training the models nor for evaluating the selected feature sets.

Subfigure (a) indicates that rhythmic descriptors do not perform well for Traditional Blues music. This is not only because the error is high in general (this can occur, for example, if a category is too hard to predict or is badly defined), but also because the error is substantially reduced if other features (e.g., audio features) are allowed to contribute to the feature sets used to train the classification models. So, the overall *multi-group feature importance* of rhythmic descriptors is low, which is indicated graphically by the large area shaded grey on the graph. As the number of all possible non-empty feature sets is typically very high ($2^N - 1$ for N features), an evolutionary algorithm is proposed in [10] to explore many different combinations of features in a non-deterministic way, guided by a process inspired by natural evolution, where feature sets can be randomly changed by *mutation* that switches feature dimensions on or off. Only fitter feature sets survive after an exit condition is fulfilled, such as a cap on the overall number of mutations.

In [10], we have also investigated the importances of different modalities and proposed the complementary concept of *multi-group feature redundancy*. However, despite a large number of experiments, the results should be treated with caution because of two limitations of the study: only random forest classifiers were used and balanced relative error was the only measure of classification quality. It was therefore unclear whether estimated importances would remain similar if other classification methods were applied or other evaluation measures used. The present study addresses both of these gaps, by investigating how stable multi-group feature performance is when classification method or evaluation measure are varied.