PARAMETER SENSITIVITY OF DEEP-FEATURE BASED EVALUATION METRICS FOR AUDIO TEXTURES

Chitralekha Gupta **Yize Wei** **Zequn Gong**
Purnima Kamath **Zhuoyao Li** **Lonce Wyse**
National University of Singapore
chitralekha@u.nus.edu, yize.wei@u.nus.edu, zequn.gong@u.nus.edu
purnima.kamath@u.nus.edu, zhuoyaoli@u.nus.edu, lonce.wyse@nus.edu.sg

ABSTRACT

Standard evaluation metrics such as the Inception score and Fréchet Audio Distance provide a general audio quality distance metric between the synthesized audio and reference clean audio. However, the sensitivity of these metrics to variations in the statistical parameters that define an audio texture is not well studied. In this work, we provide a systematic study of the sensitivity of some of the existing audio quality evaluation metrics to parameter variations in audio textures. Furthermore, we also study three more potentially parameter-sensitive metrics for audio texture synthesis, (a) a Gram matrix based distance, (b) an Accumulated Gram metric using a summarized version of the Gram matrices, and (c) a cochlear-model based statistical features metric. These metrics use deep features that summarize the statistics of any given audio texture, thus being inherently sensitive to variations in the statistical parameters that define an audio texture. We study and evaluate the sensitivity of existing standard metrics as well as Gram matrix and cochlear-model based metrics in response to control-parameter variations for audio textures across a wide range of texture and parameter types, and validate with subjective evaluation. We find that each of the metrics is sensitive to different sets of texture-parameter types. This is the first step towards investigating objective metrics for assessing parameter sensitivity in audio textures.

1. INTRODUCTION

Audio textures are rich and varied sounds that are produced by a superposition of multiple acoustic events. Unlike the sound of an individual event, such as a footstep, or the complex spectrotemporal structures and sequences of speech or music, an audio texture is defined by properties or parameters that remain constant over time [1] despite statistical variation in the sound over time or between instances. For parametric audio texture synthesis, the goal is to generate novel sounds with descriptive parameters that match those of a target texture. Standard objective evaluation metrics for audio texture synthesis include diversity and quality based measures such as the Inception score [2], and Fréchet Audio Distance (FAD) [3]. However, the existing metrics have not been studied for their sensitivity to systematic variation in the parameter values that define the synthesized audio textures.

Various audio texture synthesis algorithms have been applied to modeling a wide range of natural audio texture types including rhythmic (eg. drums, tapping), pitched (eg. windchimes, churchbells) and other natural sounds (eg. rain, wind), but with different degrees of success. Evaluation of parametric variation textures has typically been through human listening experiments [4, 5].

"Deep features" (activation patterns across layers of neural networks) have been explored for various evaluation tasks such as out-of-distribution detection in audio classification [6], perceptual image similarity evaluation [7], audio distortion assessment [3] and audio texture synthesis [1, 5, 8]. In this work we study several existing objective metrics and also explore deep features and cochlear channel model statistics for potential evaluation metrics sensitive to parameter variations. We make use of a controlled audio texture dataset [9] to study these metrics for a wide range of audio textures - rhythmic, pitched, and non-rhythmic non-pitched under systematic parametric variation. We present a comparative study of the parameter sensitivity of the metrics and validate their reliability through human listening tests.

2. RELATED WORK

2.1 Why is a parameter sensitive metric needed?

McDermott and Simoncelli [1] developed a set of statistics based on a cochlear model to describe the perceptually relevant aspects of a given audio texture. To synthesize a new audio texture, a random input is iteratively perturbed until its statistics match those of a target. This algorithm produces convincing audio for many natural textures such as insect swarms, a stream, or applause. However, human listeners give low scores on realism for resynthesized pitched and rhythmic textures, such as wind chimes, drum break, walking on gravel, and church bells. Besides the audio quality, the statistical parameters associated with these

sounds can fail to be faithfully preserved in the synthesized sounds [4]. Moments derived from the time-averaged scattering coefficients when used for resynthesis of audio textures have also shown similar performance but using fewer coefficients [10, 11]. An objective metric for predicting these human evaluations of texture synthesis failures would be valuable.

Gatys et al. [12] did seminal work on image texture synthesis that replaced hand-crafted statistics with Gram matrix statistics computed as the correlation between hidden feature activations from layers of a trained convolutional neural network (CNN). Iteratively perturbing a random input to match Gram matrices produces compelling and novel image textures. Similarly, Ulyanov et al. [8] improved the quality of synthetic textures based on the dataset from McDermott and Simoncelli [1]. Antognini et al. [5] extended Ulyanov's work by modifying the architecture with 6 parallel single-layered untrained CNNs each with a different convolutional kernel size, and computed autocorrelation and diversity losses in addition to the original Gram matrix loss. They demonstrated some improvements, but found failure modes similar to that of McDermott and Simoncelli [1] using a combination of objective and subjective evaluation. Caracalla and Roebel [13] also relied on human listening tests to show the improvement in sound quality achieved in this kind of iterative texture synthesis using a complex spectrogram audio representation.

The key take-away from these previous studies is that there is a need for evaluating the preservation of statistical parameters in synthesised audio texture but a lack of an objective metric for that purpose.

2.2 The existing audio synthesis evaluation metrics

The evaluation of generative models in terms of perceptual realism is challenging. However, various objective metrics have been formulated for assessing the quality of synthesized audio textures. For example, L2 distance, cosine distance, and signal-to-distortion ratio use a clean reference signal to compare with the modified or enhanced synthetic signal. Such signal-level metrics are agnostic to the type of audio that is being enhanced. However, these metrics may not always correlate with human perception. FAD [3] takes a different approach as a reference-independent metric that, instead of looking at individual clips, computes the distance between the distribution of embedding statistics generated on a large set of clips.

The Inception score [2, 14, 15], passes generated examples through a pre-trained classifier. The mean KL divergence between the conditional output class probabilities and the marginal distribution of the same are then calculated. The Inception score penalizes models whose examples are not easily classified into a single class, as well as models whose examples collectively belong to only a few of the possible classes. However, the Inception score is not the right tool for evaluating a model's sensitivity to parameter differences between sounds within a class.

The goal of most of these audio quality assessment metrics has been to quantify the degradation of an enhanced/modified audio signal. However, to the best of our knowledge, there is a lack of a systematic study of the sensitivity of these metrics to parameter variations in synthesized audio. Such a study is particularly challenging for audio texture synthesis because of the inherent variations that correspond to a particular parametric description.

3. METRICS

We study two standard metrics and three Gram matrix metrics for sensitivity to parametric changes to audio textures.

3.1 Existing Metrics

3.1.1 Fréchet Audio Distance (FAD)

FAD [3] is a reference-independent metric for audio quality assessment that computes the Fréchet distance [16] between the multi-variate Gaussian distributions of the embeddings of train and test set audio data. These 128 dimensional embeddings are extracted from a VGG-ish model [17] pre-trained on clean audio data for classification.

$$FAD = ||\mu_b - \mu_t||^2 + tr(\Sigma_b + \Sigma_t - 2\sqrt{\Sigma_b \cdot \Sigma_t}) \quad (1)$$

where the training and test data embeddings are assumed to have multivariate Gaussian distributions $\mathcal{N}(\mu_b, \Sigma_b)$ and $\mathcal{N}(\mu_t, \Sigma_t)$, respectively. We used the open-source model and FAD computation code [3].

3.1.2 L2 Distance

As another standard metric for our study, we compute the Euclidean distance between the two matrices X_A and X_B, representing the spectrograms of the two input audio signals to be compared, x_A and x_B respectively.

3.2 Gram matrix based Metrics

3.2.1 Gram matrix Metric (GM)

Following Ulyanov and Lebedev [8] and Antognini et al. [5], we use 2D spectrogram representations of audio for iterative updates through the CNNs used to compute a summary statistic in the form of a Gram matrix. Two audio textures sound similar if the computed Gram matrices are similar. Here, we leverage on this property of Gram matrix to develop a metric designed to be sensitive to parameter variations. We define the Gram matrix metric as the mean squared error between the Gram matrices of two textures. We hypothesize that this metric would be sensitive to parametric variations in the statistics of the same texture type such as different rates of flowing water.

Formally, we define the Gram matrix metric GM as,

$$GM = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{D}||G_A^n - G_B^n||_2^2 \quad (2)$$

where, G_A^n and G_B^n are flattened Gram matrices derived from audio clips A and B respectively for the n^{th} CNN in an ensemble of N CNNs, and D is the total number of elements in the Gram matrix. In this work, we adopted the

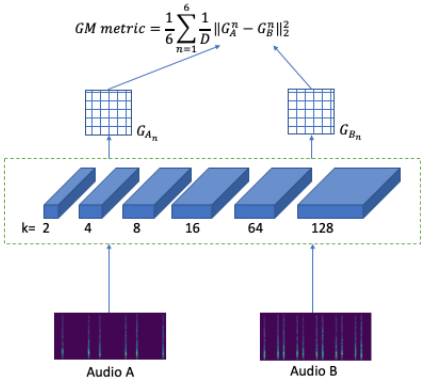

Figure 1. Gram matrix param-sensing metric computation block diagram consisting of an ensemble of six 1 layer 1d-CNNs, with different kernel sizes, k. Each CNN has 512 filters, and the weights are randomly initialized.

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

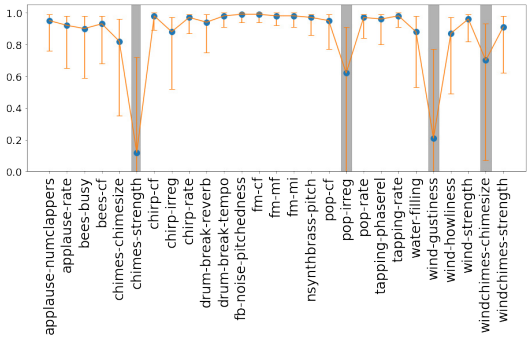

Figure 2. Correlation of perceptual distances with control parameters used to generate the textures. Grey bars indicate textures with no significant correlation.

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**

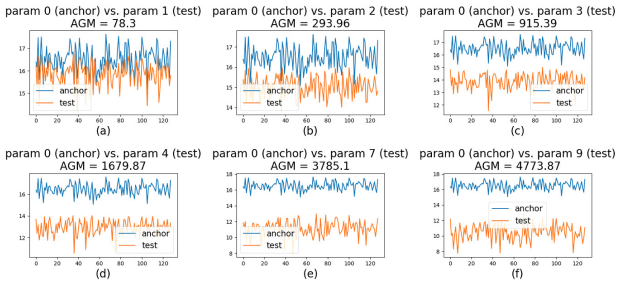

Figure 3. Gram vectors and AGM values of anchor for applause sounds with an increasing number of clappers from (a) to (f) compared to the anchor sound of a single clapper.

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

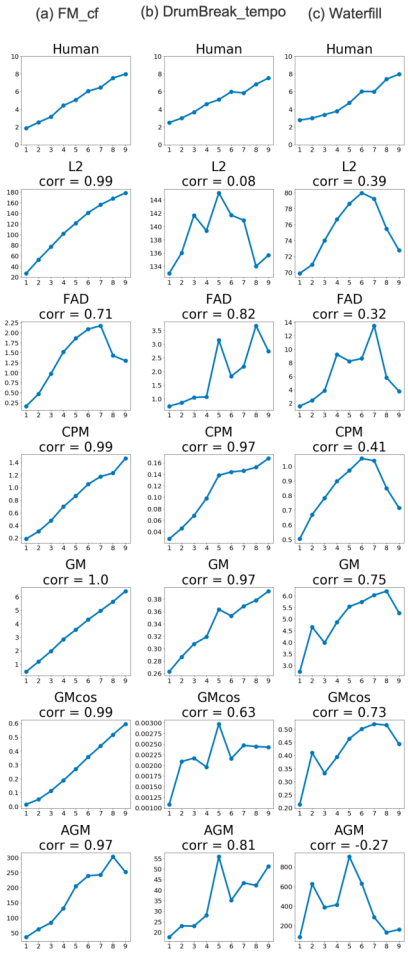

Figure 4. Trends of human responses and objective metrics to parameter variations for three texture-param combinations, (a) FM with carrier frequency variable parameter, (b) Drum-break with tempo a variable parameter, and (c) Waterfill with fill-level parameter, along with the Pearson's correlation between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