PREDICTING TIMBRE TRAITS FOR INTERPRETABLE ASSESSMENT OF MUSICAL SOUND SYNTHESIZERS

Théo CHASLE CAUCHY (theo.chasle-cauchy@ls2n.fr) [1], **Modan** [2], **Lindsey REYMORE**[3], **Fanny ROCHE**[4], and **Mathieu LAGRANGE**[5]

[1, 2, 5]**Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004**, F-44000 Nantes, France
[3]**Arizona State University, School of Music, Dance and Theatre**, Tempe, AZ, USA
[4]**Arturia**, Montbonnot Saint-Martin, France

ABSTRACT

Measuring neural audio synthesizers' performance is now routinely conducted using distribution based metrics such as the Fréchet Audio Distance (FAD). Although this metric can be correlated with human perception, it offers limited interpretability beyond ranking different approaches. In this paper, we introduce a deep neural timbre trait predictor composed of a pretrained audio neural embedding (CLAP), and a shallow learnable component. The latter is trained using the RWC musical instrument database and human judgments of 20 timbre descriptions (e.g., woody, percussive, rumbling, etc.) for 31 instruments. The resulting model shows strong correlation with average human ratings ($r = 0.66$, $p < 0.001$).

We then demonstrate the benefit of this predictor for evaluating the performance of TokenSynth, a neural sound synthesizer. First, the Mean Absolute Error (MAE) computed over the set of generated sounds under different conditioning modalities of the model provides the same ranking as a FAD computed with the RWC database as a reference, suggesting that the proposed predictors are able to provide equivalent information on a distributional basis. Second, because the model is able to qualitatively analyze isolated sounds, we can determine which generated sounds could be improved and identify specific timbral dimensions that need adjustment.

1. INTRODUCTION

The evaluation of synthesized music instruments has become increasingly important as digital audio technology continues to advance. Whether for virtual instruments or AI-generated sounds, the ability to objectively measure how closely a synthesized sample matches its real-world counterpart is crucial for both developers and musicians. The standard for automatic assessment is the Fréchet Audio Distance (FAD) [1]. Originally inspired by the Fréchet Inception Distance used in image generation, FAD adapts this concept to the audio domain by comparing the statistical distributions of real and synthesized audio features. It provides a quantitative way to assess the perceptual quality and authenticity of synthesized instrument samples. To produce a meaningful metric, it requires embeddings that are trained on specific audio datasets—such as music or environmental sounds [2, 3].

Although FAD offers a valuable means of ranking the overall quality of synthesized samples, it remains a global metric with limited interpretability. Specifically, it does not provide actionable insights into which perceptual attributes might be responsible for discrepancies between real and synthesized sounds. This limitation underscores the need for more detailed and interpretable metrics, to our knowledge absent in the state-of-the-art, especially those related to timbre, as it is an essential dimension of audio perception that is difficult to formalize.

Timbre has been the subject of extensive research in audio signal processing and music information retrieval. Because it requires bridging the gap between human perception of timbre and machine listening, timbre qualification remains challenging. Starting with pioneering work by Plomp (1970) [4], Wessel (1973) [5], and Grey (1977) [6], multidimensional scaling (MDS)—which uses pairwise dissimilarity ratings to model perceptual relationships among sounds—has become a common approach to mapping timbre space. MDS does not offer explanations of the resulting dimensions, which must instead be interpreted by the researchers. These dimensions are typically interpreted in terms of acoustic features; to the best of our knowledge, McAdams et al. (1995) [7] were the first to demonstrate that certain perceptual descriptors are correlated with the dimensions of the timbre space introduced in their study. Peeters et al. [8] proposed descriptors based on signal processing (e.g., spectral centroid, attack time) integrated into the MPEG-7 [8] audio description scheme for standardized timbre representation. Later, the Audio Commons initiative advanced the field by developing perceptual models linking signal processing metrics (e.g., zero crossing rate, spectral centroid, energy ratios) to high-level perceptual attributes (e.g., roughness, brightness, sharpness, warmth) [9–11], offering a more interpretable timbre description, particularly for non-musical sounds. The Timbre Toolbox [12, 13] emerged as a comprehensive resource, employing advanced techniques such as the Short-Time Fourier Transform and Constant-Q Transform to extract over 100 descriptors—including spectral, temporal, harmonic, and modulation features. Some of these

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

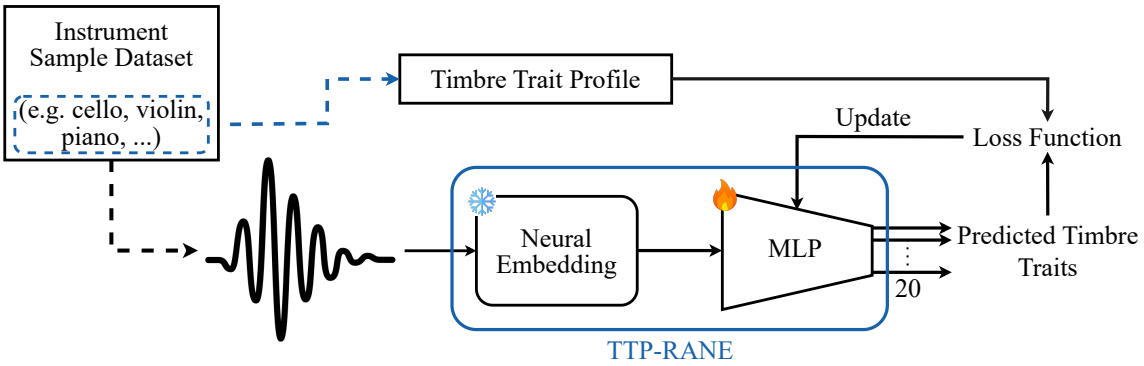


Figure 1: Timbre Traits Profiles - Reweighted Audio Neural Embedding (TTP-RANE): First, the frozen neural embedding model computes each audio sample embedding. Second, the trained MLP assigns values for the 20 timbre traits. During training, the loss function between predictions and ground truth (Timbre Trait Profiles) is computed to update the weights of the MLP. Solid lines indicates the process of the approach, dashed lines indicates where the dataset information is used.

acoustic features, like spectral centroid (linked to perceived brightness) and attack time (distinguishing percussive from sustained sounds), correlate with perceptual attributes [14, 15]; however, relationships between acoustic correlates and semantic labels are often complex, with some research demonstrating significant advantages for nonlinear approaches [16, 17].

Deep learning approaches, such as those by Pons et al. [18], introduced Convolutional Neural Networks (CNNs) to learn timbre representations directly from log-mel spectrograms, achieving state-of-the-art performance in tasks like instrument recognition and music auto-tagging. However, these features lack interpretability.

In this paper, we address the lack of interpretable metrics for sound synthesizer assessment by introducing a predictor based on a 20-dimensional model of "musical instrument timbre qualia" proposed by Reymore and Huron [19].

Specifically, we leverage the data collected in this study to train a model to predict values across the 20 dimensions (called timbre traits) from audio input. This multi-layer-perceptron model is trained on deep neural embeddings of instrument samples from the RWC dataset, comprising 31 of the 34 instruments rated in a 2021 study by Reymore [20].

We then demonstrate the benefit of this predictor for evaluating the performance of TokenSynth [21], a sound synthesizer. First, we synthesize three sets of samples using three different conditionings of the model. We compute the predictions' Mean Absolute Error (MAE) on each of the synthesized sample set and demonstrate that it provides the same ranking as a FAD computed with the RWC database as a reference, suggesting that the proposed predictors are able to provide equivalent information on a distributional basis. Second, because the model is able to analyze sounds one by one, we can use it to analyze the timbral dimensions of generated sounds to improve the generation.

The contributions of this paper are thus two-fold: 1) we introduce automatic timbre trait prediction as a machine listening task and demonstrate its feasibility; 2) we demonstrate the potential of our TTP prediction method "Reweighted Audio Neural Embedding" (TTP-RANE) to complement FAD in order to obtain qualitative guidance for the evaluation of generative instrument audio synthesis.

1. sparkling/brilliant	11. hollow
2. focused/compact	12. woody
3. pure/clear	13. airy/breathy
4. open	14. nasal/reedy
5. ringing/long decay	15. raspy/grainy
6. resonant/vibrant	16. rumbling/low
7. sustained/even	17. direct/loud
8. soft/singing	18. percussive
9. watery/fluid	19. shrill/noisy
10. muted/veiled	20. brassy/metallic

Table 1: Dimensions of the Timbre Trait Profiles (TTP).

The paper is organized as follows. Section 2.1 introduces the work of Reymore and what timbre traits are. Section 2 presents the training procedure and model performance. Section 3 presents the application of the model to evaluating synthesized sounds. Code and audio examples are made available. [1]

2. PREDICTING TIMBRE TRAITS

2.1 Definition

The 20 dimensions of the timbre qualia model proposed by Reymore and Huron [19] were empirically derived through a two-stage process. First, open-ended interviews with 23 musicians generated 77 descriptive categories of timbre qualia, which were refined via content analysis and pile sorting. Second, 460 musicians rated these categories for 20 instruments. Principal Component Analysis (PCA) was applied to the ratings to reduce dimensionality, and feedback from musicians on component relevance was integrated, resulting in the 20 dimensions of the model. Reymore and Huron refer to the dimensions as timbre traits; these traits are presented in Table 1.

In a subsequent study, Reymore [20] collected rating data from 243 musicians on each of the 20 dimensions in order

[1] Companion page: `https://theochaslecauchy.github.io/paperTTPSynthesizerAssessment/`

to generate Timbre Trait Profiles (TTPs) for 34 common orchestral and wind ensemble instruments. As with the interviews and rating tasks from Reymore and Huron [19], ratings used to create the TTPs were based on participants' imagined archetypical instrument sounds. For each instrument rated, participants were first asked to auralize [2] a typical sound played on that instrument, then to rate the imagined sound in the range $[\![1, 7]\!]$ on each dimension of the timbre model. Each TTP includes the average participant ratings on 20 dimensions.

The TTPs presented by Reymore in [20] were generated by averaging human ratings for each instrument. In this study, we present a TTP prediction method called TTP-RANE illustrated in Figure 1. We trained a multi-layer perceptron model to predict TTPs from deep neural embeddings of instrument samples. To this end, we used the RWC audio dataset, with the TTPs assessed by Reymore [20] assigned to each sample according to its instrument class.

2.2 Dataset

The RWC dataset [22] contains 91.6 hours of instrument samples, including 31 of the 34 musical instruments rated by Reymore [20], listed in Figure 2 [3].

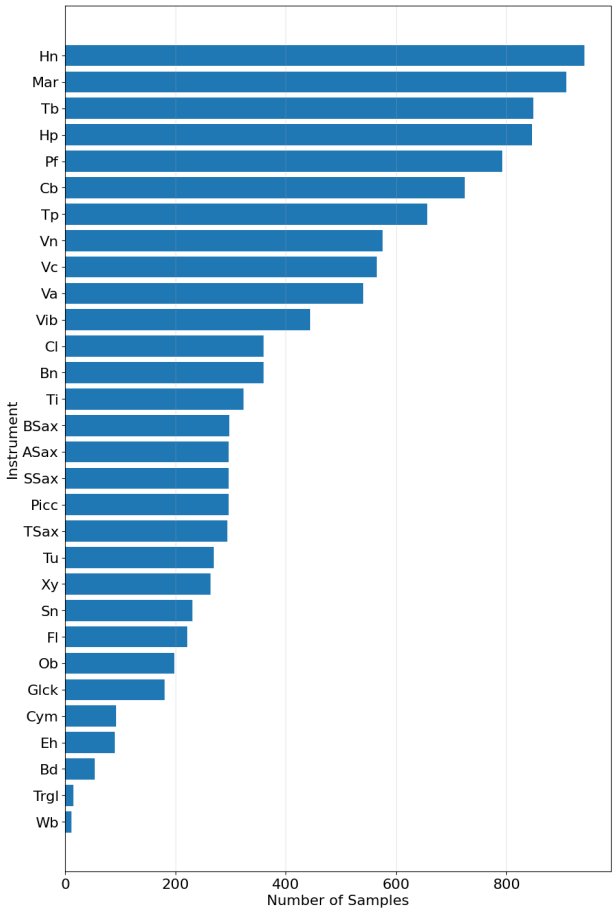


Figure 2: Musical instruments and their number of samples from RWC dataset.

[2] "imagine the sound of"

[3] Abbreviations correspondences can be found on the companion page.

For pitched instruments, each recording contains every note in the instrument's pitch range, played successively. For percussive instruments (bass drum, cymbal, wood block, snare drum), which lack a defined pitch range, the recordings consist of several strikes or variations in playing style. Most instruments have recordings from different instrument brands or sub-instrument categories (e.g., for example for the piano, Pianoforte Yamaha, Steinway and Bosendorfer can be found). To obtain isolated note samples, each recording is segmented into individual notes. We first normalize samples and detected single notes with an initial threshold of -70 dB, decreasing it by 5dB successively if one of the resulting split audio files lasts more than 15 seconds. Only samples in their archetypical playing styles (ordinario) are considered for this study, in order to avoid playing styles that would diverge from the typical TTP of an instrument. The final number of extracted samples for each instrument is shown in Table 2, corresponding to a total of 11 hours of audio. The average duration of the samples is 3.22 seconds and the standard deviations 1.4 seconds.

2.3 Neural Embeddings

We consider and compare predictors built upon 4 deep neural embeddings: VGGish [23], the Music undERstanding model with large-scale self-supervised Training (MERT) [24], Laion Contrastive Language-Audio Pretraining (CLAP) [25] and its music-only version (CLAP-Music).

VGGish [23] is an audio classifier trained on a subset of a large audio dataset extracted from YouTube videos, called YouTube-100M, which contains 350,000h of audio data with video-level class labels. VGGish adopts a CNN-based approach to extract features from log-mel spectrograms with 64 frequency bins and 10-ms hops to compute audio embeddings.

MERT [24] is an embedding specialized in music analysis through a self-supervised learning framework using teacher-student methods: a combination of teachers, including an acoustic teacher based on Residual Vector Quantization - Variational AutoEncoder, a musical teacher based on the Constant-Q Transform, and a BERT-style transformer encoder as student model. The MERT models were trained on an extensive in-house private music dataset comprising 160,000 hours of audio data. Among these, the "MERT-v1-95M" model stands out due to its performance on music understanding benchmarks.

CLAP is trained to learn multi-modal representations using both an audio and a text encoder. First introduced in [26] by Microsoft, LAION later presented a CLAP model with significantly better performance by training it on the large-scale LAION-Audio-630k and AudioSet datasets, enhanced with keyword-to-caption augmentation. Their best general-purpose model, "630k-audioset-best", is trained on approximately 10,000 hours of audio and achieves state-of-the-art results on standard audio-text retrieval benchmarks, making it the most effective model for broad audio understanding.

We also considered LAION CLAP models fine-tuned on

music and speech data. Among these, the "music audioset epoch 15 esc 90.14" model achieved superior performance on music-specific benchmark tasks. We hereafter refer to this model as "CLAP-Music".

We compute the CLAP, CLAP-Music, Mert and VGGish embeddings using the "fadtk" python package considered by Gui et al. [27].

2.4 Supervised approach

We present here a TTP prediction supervised approach called TTP-RANE. This method, illustrated in Figure 1, is composed of a frozen deep neural embedding following by a shallow learnable MLP reweighting the neural embedding. "Reweighting" refers here to the process of using the MLP to recalibrate the weights of the neural embedding features to achieve this prediction task. We train this MLP to predict TTPs from audio embeddings of RWC dataset samples.

To this end, the embeddings dataset is shuffled and split into a training and a validation set (80%-20%). We use the TTP as ground truth labels. The values of timbre traits [20] belong to the range $[\![1, 7]\!]$. First, the values are normalized to the interval [0,1], and a sigmoid is applied to the output values of the model to predict values in this interval. We used the Mean Square Error (MSE), commonly applied in regression problems, as our loss function.

Given the limited amount of ground truth data, we evaluate the models' performance using a cross-evaluation approach based on instrument. That is, for each instrument, we train a model using samples from all the other instruments and then evaluate its performance on samples from the target instrument. To evaluate a given model, we thus train 31 separate models, one for each instrument, with the same architecture and training conditions, and we evaluate each model on all its unknown instrument samples (training set and validation set).

We test 3 MLP architectures: 1) one with no hidden layer, 2) one with one hidden layer of size 256, and 3) a funnel-shaped architecture with 2 hidden layers of sizes 256 and 128. We use a learning rate of 0.01 scheduled with a "Reduce on Plateau" scheduler, an Adam optimizer and a maximum of 300 epochs per training. We apply early stopping on the validation set with a patience of 20 epochs; that is, the training stops when the loss of the predictions on the validation set does not decrease during 20 consecutive epochs, and we keep the model with the lowest validation loss.

2.5 Unsupervised approach

To measure to benefit of using TTP-RANE, we consider an unsupervised approach that does not consider any specialized prior knowledge on musical timbre, called "Text-to-Audio Similarity" (T2ASim). T2ASim predicts timbre trait values by leveraging the similarity between audio and text embeddings in the CLAP space. Specifically, for an audio sample with a CLAP embedding $\mathbf{E_i}$ and a text embedding $\mathbf{t}_j$ representing the j-th trait, the prediction process is as follows:

	Embedding	Pearson ↑
T2ASim	CLAP [25]	.101*
TTP-RANE	MERT [24]	.578*
	VGGish [23]	.581*
	CLAP-Mus. [25]	.631*
	CLAP [25]	**.663***
Hum. Ratings		.698*

Table 2: Comparison of Pearson correlations of the predicted timbre traits values by models and human ratings (* p value under 0.001).

1. Embeddings Distance: Computation of the distance $d(\mathbf{E_i}, \mathbf{t}_j)$ between the audio embedding $\mathbf{E_i}$ and the trait's text embedding $\mathbf{t}_j$.

2. Normalization: Divide the distance by $\max_{i,j}(d(\mathbf{E_i}, \mathbf{t}_j))$.

3. Prediction: Derive the predicted trait value as $1 - d(\mathbf{E_i}, \mathbf{t}_j)$.

2.6 Results

To assess the performance of the evaluated models, we consider the Pearson correlation between the predicted TTPs and the ground truth TTPs. Each element of the prediction array is the average of the predictions for a specific timbre trait and instrument in the RWC dataset. As there are 20 timbre traits and 31 instruments, this results in 620 points per model for the correlation computation. The higher the Pearson correlation values, the better the timbre trait values prediction.

Cross-evaluation results are displayed in Table 2. For each embedding type, Table 2 presents only the model with no hidden layer, as we found it to be the architecture with the best cross-evaluation performance. We observe on the training loss that the models with 1 or more hidden layers overfit on the training set resulting in lower performances during the cross-evaluation, contrary to the model with no hidden layer that generalizes better to instruments not included in the training set.

For the human ratings, the Pearson correlation is computed between the ratings collected by Reymore [20] and the ground truth values (that correspond to the average human ratings). This gives an estimate of the inter-rater agreement of human ratings. [4]

The model achieving the best cross-evaluation performance is the model trained on CLAP embeddings. Thus for the assessment of the synthesizer, we train a new model on all 31 instruments using CLAP embeddings and the same parameters as previously, referring to it hereafter as TTP-RANE-CLAP.

[4] Inter-rater agreement refers to the degree of consistency or concordance among ratings provided by different human evaluators.

3. ASSESSING TOKENSYNTH

In this section, we demonstrate the potential of using the timbre trait predictor to assess the performance of a musical instrument synthesizer: TokenSynth.

3.1 Overall Synthesis Assessment

TokenSynth [21] is a neural synthesizer that leverages token-based audio representations to enable text-to-instrument synthesis. This synthesizer takes as input a MIDI token and a CLAP embedding. It employs a decoder-only transformer that computes audio tokens from inputs, decoded by a pre-trained Descript Audio Codec (DAC) to generate audio samples. The model was trained on a dataset combining NSynth [28] and Lakh MIDI [29], totaling over 9.53 million samples, which was further augmented with digital audio effects to enhance timbral diversity.

For each instrument in Table 2, we randomly sample 100 pitches in the corresponding pitch range of the instrument recordings in the RWC dataset (e.g., G3 to E7 for the violin). For each instrument, we sample pitches from a Gaussian distribution whose mean is set at the midpoint of the instrument's pitch range and whose standard deviation is set to 20% of the total pitch range. CLAP embeddings enable the projection of text and audio in the same space. As TokenSynth synthesis is conditioned with a CLAP embedding, we can use text and/or audio conditioning. Therefore, for each sampled pitch of an instrument **I**, we synthesize 3 sounds with different embeddings as input: 1) Text conditioning: the text embedding of the instrument **I** name (e.g., "CELLO", "SNARE DRUM"); 2) Audio conditioning: the averaged audio embedding of all samples of the instrument **I** from the RWC dataset; 3) Text-Audio conditioning: the mean embedding of both previous embeddings.

To assess the synthesizer, we first compute FAD between RWC samples and TokenSynth samples, then we compute the MAE between predicted TTPs on synthesized samples and the ground truth. For the sake of comparison, we also compute the MAE between predicted TTPs on the RWC samples (training set and validation set) and the ground truth.

Results are shown in Table 3. We observe that there are no significant statistical difference between different conditioning generations with respect to their MAE. Furthermore, MAE between predicted TTPs and the ground truth gives the same ranking as the FAD, suggesting that TTP-RANE-CLAP provides, at a macro scale, equivalent information about the performance of the studied synthesizer. We can now go deeper in the analysis. In the following, we focus on the audio conditioned synthesis.

3.2 Wood Block Synthesis Assessment

We can start the analysis of this synthesis by observing the predicted TTP of the instrument with the highest MAE, here, the wood block (MAE of 0.24). This indicates that the wood block is the instrument which synthesized samples have the furthest predicted TTPs from the ground truth TTP. The wood block is the less represented instrument in the RWC dataset, this could lead to less accurate predictions by TTP-RANE. However, the investigation on synthesized wood block samples confirms these results. Figure 3 shows the predicted TTP of the synthesized wood block, that is, the average predicted TTP across all samples.

	FAD ↓	MAE ↓
RWC samples		.058 ± .056
Text conditioned samples	.57	.180 ± .128
Audio conditioned samples	.53	**.172 ± .123**
Text-Audio cond. samples	.54	.173 ± .124

Table 3: FAD between RWC and synthesized samples and MAE of the predicted timbre trait values on the synthesized samples compared to the MAE on the RWC dataset.

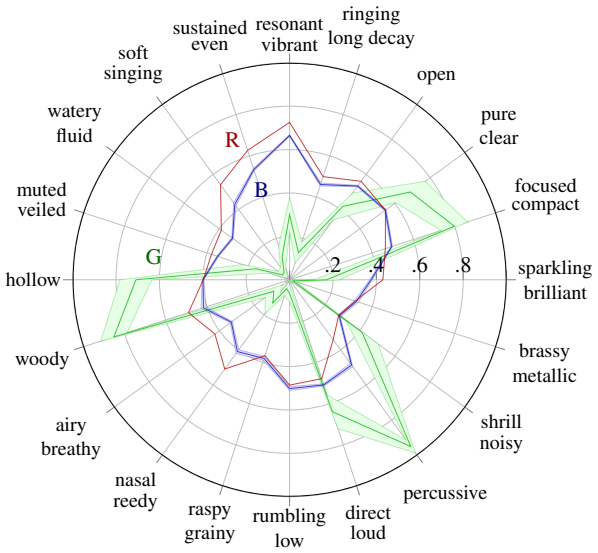


Figure 3: Radar plots of TTP for Wood Block with their respective 95% confidence interval for Reymore's results (Green); Averaged predicted TTP on synthesized wood block samples (Blue); Predicted TTP of the sample with highest difference between "percussive" prediction and ground truth (Red).

We observe on the wood block averaged predicted TTP that the timbre traits with the highest errors are "woody" and "percussive." While "woody" can not be easily assessed on a spectrogram by a non-specialist, we can make observations with respect to the "percussive" trait. Figure 4 shows the spectrograms of a) the synthesized wood block sample with the highest difference between predicted "percussive" value and ground truth, and b) the RWC wood block sample with the lowest difference between predicted "percussive" value and ground truth.

The wood block is a percussive instrument, so its ground truth value of the "percussive" trait is about 0.93. The prediction of this trait for the synthesized sample is 0.32. We can observe on the spectrograms that the synthesized sample is sustained, contrary to the RWC sample, which decays quickly after impact. This explains why its "percus-

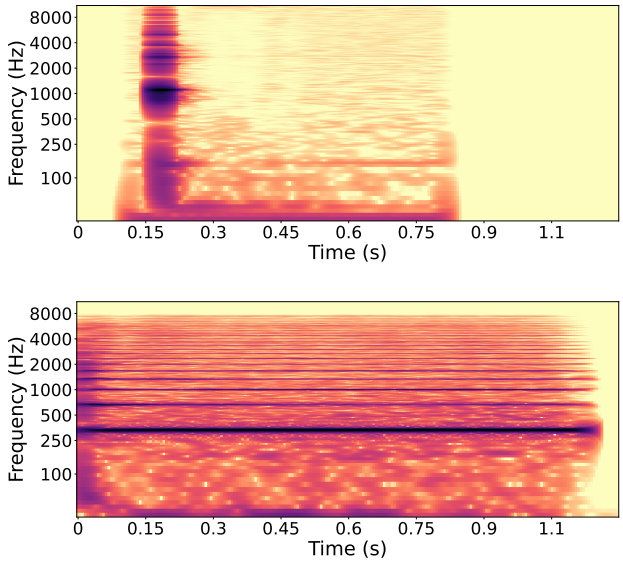


Figure 4: Above: Spectrogram of the RWC wood block sample with the lowest difference between predicted "percussive" value and ground truth. Below: Spectrogram of the synthesized wood block sample with the highest difference between predicted "percussive" value and ground truth

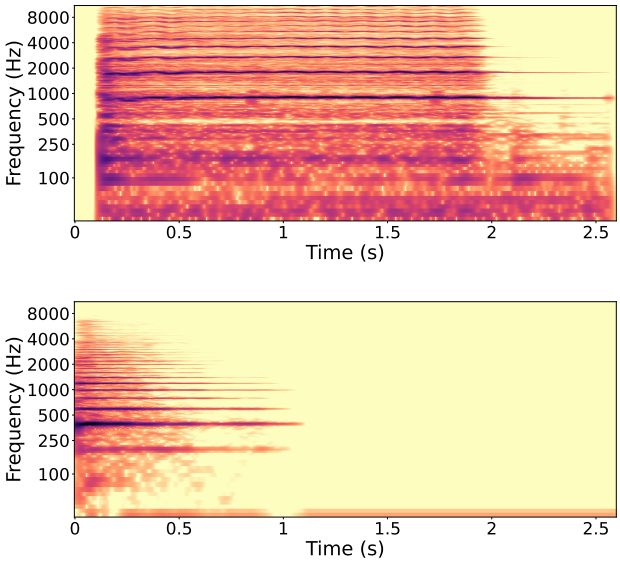


Figure 6: Above: Spectrogram of the RWC cello sample with the lowest difference between predicted "resonant-vibrant" value and ground truth. Below: Spectrogram of the synthesized cello sample with the highest difference between predicted "resonant-vibrant" value and ground truth.

sive" trait prediction is far below the ground truth.

3.3 Cello Synthesis Assessment

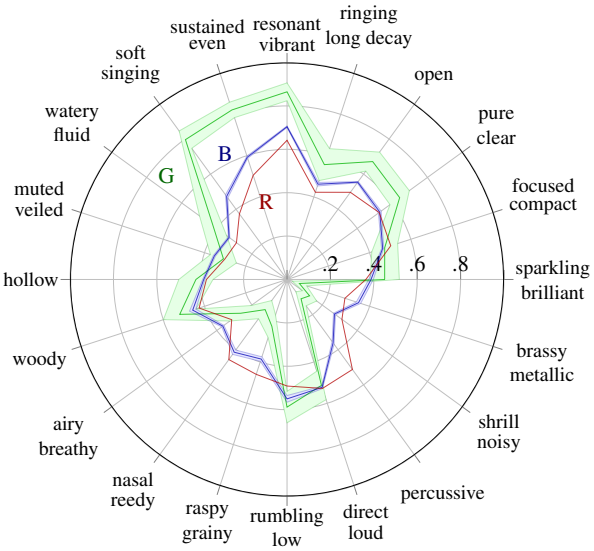


Figure 5: Radar plots of TTP for Cello with their respective 95% confidence interval for Reymore's results (Green); Averaged predicted TTP on synthesized cello samples (Blue); Predicted TTP of the sample with highest difference between "resonant-vibrant" prediction and ground truth (Red).

We can also analyze more subtle defects by observing any timbre trait of any instrument. For instance, we can analyze the "resonant/vibrant" trait of the cello. Its predicted TTP is presented in Figure 5. The "resonant/vibrant" ground truth value is 0.865, the highest value for this trait in the 31 instruments considered here, suggesting it is an essential timbral quality for the cello. Figure 6 shows the spectrograms of a) the synthesized cello sample with the highest difference between predicted "resonant-vibrant" value and ground truth, and b) the RWC cello sample with the lowest difference between predicted "resonant/vibrant" value and ground truth. The prediction of this trait for the synthesized sample is 0.66. By observing the rapid amplitude decay of the synthesized sample's spectrogram and by listening to it, we notice that TokenSynth synthesized a plucked note sample whereas the RWC sample uses bowing technique. Difference in performance technique leads of course to significant timbre differences. The spectrogram of the RWC sample shows harmonics whose frequencies fluctuate slightly over time, plausibly creating a "vibrant" perceptual quality related to the pronounced vibrato in the sample. In contrast, the synthesized sample's spectrogram lacks these frequency oscillations in its harmonics. This difference helps explain why the synthesized sample receives a lower prediction score for the "resonant/vibrant" attribute. [5]

Following this method, we can utilize the predicted TTPs to assess the synthesized instrument samples, identify shortcomings, and suggest new design strategies to potentially improve the performance of the synthesizer.

4. DISCUSSION

In this paper, we propose a new machine listening task: Timbre Traits Profiles prediction. We 1) demonstrate the

[5] The sounds discussed here as well as others can be listened to via the companion page.

effectiveness of a simple learnable reweighting of deep neural audio embeddings and 2) exemplify the use of TTP prediction for assessing the performance of the musical sound synthesizer TokenSynth.

The objective is to predict the timbres of instrument samples using TTP from Reymore [20], which is constructed from human ratings, as ground truth. We introduce a deep neural timbre trait predictor called TTP-RANE composed of a pre-trained neural embedding and a shallow multilayer perceptron, reweighting the neural embedding. To train our model, we compare four deep neural embeddings: CLAP (and its music-only version), MERT, and VGGish on samples from the RWC dataset. TTP-RANE performance is evaluated via Pearson Correlation during cross-evaluation, with the best result (0.66) achieved by a no-hidden-layer model trained on CLAP embeddings. We then train a global model called TTP-RANE-CLAP on all 31 instruments and apply it to assess the synthesis quality of TokenSynth, guided by CLAP embeddings, across three synthesis types: text-conditioned, audio-conditioned, and a hybrid conditioned by averaging text-audio embeddings. Audio-conditioned synthesis yields the most accurate TTPs predictions, and we observe that text conditioning added to audio conditioning does not improve the predictions. The ranking of synthesis quality based on the Mean Absolute Error (MAE) between predictions and ground truth aligns with the Fréchet Audio Distance (FAD) using the RWC dataset as a reference, indicating the TTP-RANE robustness in capturing distributional timbre information. Finally, we demonstrate how TTP prediction can be utilized to analyze and interpret synthesis defects, using the wood block and the cello as examples.

From a technical point of view, future work will address 1) new deep neural architectures to improve our TTP prediction model and 2) new ways to condition or train musical audio synthesizers using respectively a TTP-RANE inspired conditioner or loss. Finally, while the TTP-RANE-CLAP model has been studied here for the sole evaluation of musical audio synthesizer, it may also be useful in music analysis. Reymore [17] suggests that TTPs may be used to identify when instruments are utilized in archetypical v. atypical or unexpected ways. TTP-RANE offers a way of accomplishing this computationally, by predicting timbre profiles for notes or excerpts from a recorded piece, and using MAE as a metric for typicality. This metric could be used in music analysis, both to support close readings and to test hypotheses through corpus studies. Additionally, a refined TTP prediction model could inform understanding of timbre semantics, potentially contributing to characterizing the nature of timbral interactions with other musical parameters, such as pitch, loudness, and articulation. Initial work has been done in this area [30, 31], but the enormous variation possible across musical parameters presents a significant challenge, as it cannot be practically modeled through human judgments. TTP prediction could be thus used to map an instrument's timbral affordances across registers, dynamic ranges, and different playing techniques. Further testing could also determine whether TTP prediction is a promising approach to quantifying timbral qualities of instrumental blends. These timbral affordance maps would have practical applications in analysis, composition, and orchestration by offering access to average perceptual qualities of particular timbres without needing to acquire new sets of human judgments for each specific context.

5. REFERENCES

[1] K. Kilgour, M. Zuluaga, D. Roblek, and al, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," in *INTERSPEECH*, 2019.

[2] M. Tailleur, J. Lee, M. Lagrange, and al, "Correlation of fréchet audio distance with human perception of environmental audio is embedding dependent," in *EUSIPCO*, 2024.

[3] A. Gui, H. Gamper, S. Braun, and al, "Adapting frechet audio distance for generative music evaluation," in *ICASSP*, 2024.

[4] R. Plomp, "Timbre as a multidimensional attribute of complex tones," in *Frequency Analysis and Periodicity Detection in Hearing*. Sijthoff, 1970.

[5] D. L. Wessel, "Psychoacoustics and music: A report from michigan state university," *PACE: Bulletin of the Computer Arts Society*, 1973.

[6] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *The Journal of the Acoustical Society of America*, 1977.

[7] S. McAdams, S. Winsberg, S. Donnadieu, and al, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychological Research*, 1995.

[8] G. Peeters, S. McAdams, and P. Herrera, "Instrument sound description in the context of MPEG-7," in *ICMC*, 2000.

[9] A. Pearce, S. Safavi, T. Brookes, and al, "Release of timbral characterisation tools for semantically annotating non-musical content," AudioCommons, Report D5.8, 2019.

[10] A. Pearce, T. Brookes, and R. Mason, "First prototype of timbral characterisation tools for semantically annotating non-musical content," AudioCommons, Report D5.2, 2016.

[11] A. Pearce, S. Safavi, T. Brookes, and al, "Second prototype of timbral characterisation tools for semantically annotating non-musical content," AudioCommons, Report D5.6, 2017.

[12] G. Peeters, B. Giordano, P. Susini, N. Misdariis, and al, "The timbre toolbox: Audio descriptors of musical signals," *Journal of the Acoustical Society of America*, 2011.

[13] S. Kazazis and S. McAdams, "Timbre toolbox-r2021a," 2023. [Online]. Available: https://github.com/MPCL-McGill/TimbreToolbox-R2021a

[14] E. Schubert and J. Wolfe, "Does timbral brightness scale with frequency and spectral centroid," *Acustica*, 2006.

[15] M. Caetano, C. Saitis, and K. Siedenburg, *Audio Content Descriptors of Timbre*, 2019.

[16] S. McAdams, C. Douglas, and N. N. Vempala, "Perception and modeling of affective qualities of musical instrument sounds across pitch registers," *Frontiers in Psychology*, 2017.

[17] L. Reymore, E. Beauvais-Lacasse, B. K. Smith, and al, "Modeling noise-related timbre semantic categories of orchestral instrument sounds with audio features, pitch register, and instrument family," *Frontiers in Psychology*, 2022.

[18] J. Pons, O. Slizovskaia, R. Gong, and al, "Timbre analysis of music audio signals with convolutional neural networks," in *EUSIPCO*, 2017.

[19] L. Reymore and D. Huron, "Using auditory imagery tasks to map the cognitive linguistic dimensions of musical instrument timbre qualia," *Psychomusicology: Music, Mind, and Brain*, 2020.

[20] L. Reymore, "Characterizing prototypical musical instrument timbres with timbre trait profiles," *Musicae Scientiae*, 2021.

[21] K. Kim, J. Koo, S. Lee, and al, "Tokensynth: A token-based neural synthesizer for instrument cloning and text-to-instrument," *CoRR*, 2025.

[22] M. Goto, H. Hashiguchi, T. Nishimura, and al, "Rwc music database: Music genre database and musical instrument sound database," in *ISMIR*, 2003. [Online]. Available: https://doi.org/10.5281/zenodo.1415536

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ITAIMLE*, 2015.

[24] Y. Li, R. Yuan, G. Zhang, and al, "Mert: Acoustic music understanding model with large-scale self-supervised training," in *ICLR*, 2024.

[25] Y. Wu, K. Chen, T. Zhang, and al, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP*, 2024.

[26] B. Elizalde, S. Deshmukh, M. A. Ismail, and al, "Clap: Learning audio concepts from natural language supervision," *CoRR*, 2022.

[27] A. Gui, H. Gamper, S. Braun, and al, "Adapting frechet audio distance for generative music evaluation," in *ICASSP*, 2024.

[28] J. Engel, C. Resnick, A. Roberts, and al, "Neural audio synthesis of musical notes with wavenet autoencoders," *JML*, 2017.

[29] C. Raffel, "Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching," Ph.D. dissertation, 2016. [Online]. Available: https://doi.org/10.7916/D8N58MHV

[30] L. Reymore, J. Noble, C. Saitis, C. Traube, and al, "Timbre semantic associations vary both between and within instruments: An empirical study incorporating register and pitch height," *Music Perception: An Interdisciplinary Journal*, 2023.

[31] L. Reymore, "Variations in timbre qualia with register and dynamics in the oboe and french horn," *Empirical Musicology Review*, 2021.