UNVEILING THE TIMBRE LANDSCAPE: A LAYERED ANALYSIS OF TENOR SAXOPHONE IN RAVE MODELS

Nádia CARVALHO (nscarvalho@fe.up.pt) (0000-0001-6882-5138)[1],
Jorge SOUSA (jorgevidalsousa@ua.pt) (0009-0004-5920-2448)[2],
Gilberto BERNARDES (gba@fe.up.pt) (0000-0003-3884-2687)[1], and
Henrique PORTOVEDO (henriqueportovedo@ua.pt) (0000-0003-1258-5459)[2]

[1]**University of Porto – Faculty of Engineering and INESC TEC**, Porto, Portugal
[2]**University of Aveiro and INET-md**, Aveiro, Portugal

ABSTRACT

This paper presents a comprehensive investigation into the explainability and creative affordances derived from navigating a latent space generated by Realtime Audio Variational AutoEncoder (RAVE) models. We delve into the intricate layers of the RAVE model's encoder and decoder outputs by leveraging a novel timbre latent space that captures micro-timbral variations from a wide range of saxophone extended techniques. Our analysis dissects each layer's output independently, shedding light on the distinct transformations and representations occurring at different stages of the encoding and decoding processes and their sensitivity to a spectrum of low-to-high-level musical attributes. Remarkably, our findings reveal consistent patterns across various models, with the first layer consistently capturing changes in dynamics while remaining insensitive to pitch or register alterations. By meticulously examining and comparing layer outputs, we elucidate the underlying mechanisms governing saxophone timbre representation within the RAVE framework. These insights not only deepen our understanding of neural network behavior but also offer valuable contributions to the broader fields of music informatics and audio signal processing, ultimately enhancing the degree of transparency and control in co-creative practices within deep learning music frameworks.

1. INTRODUCTION

The intersection of music, deep learning, and creative exploration has spurred transformative advancements in understanding and manipulating sound, offering tools for creating immersive soundscapes [1], transforming existing music [2–4], and synthesizing entirely new sounds [5–7].

WaveNet [8] and SampleRNN [9] were among the pioneering deep-learning approaches for audio modeling in its raw waveform, but their reliance on extensive data and parameters often led to slow synthesis and error accumulation due to their autoregressive nature. Building upon WaveNet, Engel et al. introduced NSynth [5] to tackle representation learning. Additionally, Kumar et al. [10] proposed leveraging generative adversarial modeling to address parallel audio modeling. In the realm of real-time audio processing tasks, a significant milestone is reached with Realtime Audio Variational AutoEncoder (RAVE) models [4]. RAVE models leverage variational autoencoders' capacity to learn audio data representations, enabling dynamic manipulation and synthesis in real-time, and the integration of AI-driven sound design into real-time production workflows [11], such as live performances [12–14] to interactive installations [15].

While these models excel at learning the structures of the audio data, they often remain opaque to users, hindering interaction and control [16]. Navigating RAVE's latent space proves challenging due to the complex and highly-dimensional nature of sound. Nevertheless, transparent representations in audio latent spaces are indispensable for advancing music generation. They offer insight into how the model interprets musical timbres, enabling users to refine outputs and fostering collaboration and innovation. A transparent latent space encourages diverse contributions to model development.

Moreover, direct interaction with musical attributes enhances user engagement and customization. Ethical concerns underscore the need for transparency to manage risks and ensure accountability. Efforts to improve the accessibility of latent space representations encompass various strategies. Intuitive visualization tools and detailed documentation elucidate the space's structure and meaning, empowering users to navigate it effectively. Democratizing access to transparent latent space representations unleashes the full potential of generative models like RAVE, nurturing creativity, collaboration, and innovation in music composition and synthesis.

This paper undertakes a thorough investigation of RAVE models with the objective of elucidating the characteristics of their latent spaces. A particular focus is directed towards a specialized timbre latent space (timbre is understood as all auditory sensations other than pitch, loudness, and perceived duration [17]), meticulously designed to capture the intricate nuances inherent in saxophone extended techniques. We explore the layers of RAVE's encoder and decoder outputs, revealing the transformations and representations at each level through a dual evalua-

tion approach: 1) a perceptual subjective assessment involving a detailed analysis of the outputs, layer by layer, and comparison across different models trained on diverse datasets; and 2) an objective examination of the extent to which the models' encoding's latent spaces recognize semantically meaningful timbre attributes. This examination is conducted by quantifying intra- and inter-segment distances per technique, register, dynamic, and their combinations across all sounds in the provided dataset. Our evaluation aims to elucidate the fundamental mechanisms governing tenor saxophone timbre representation within the RAVE framework.

Our paper is organized as follows. Section 2 introduces the pre-trained models used for comparison and the datasets employed for their training. Additionally, we outline our methodology for curating a comprehensive collection of sounds representing the entirety of the tenor saxophone dataset, encapsulating the timbre latent space. Section 3 delineates the methodology utilized to evaluate the transparency derived from navigating a latent space generated by RAVE models. This section offers detailed insights into the twofold experimental setup. Subsequently, Section 4 presents and discusses the results obtained. Finally, Section 5 summarizes the conclusions drawn from our study and proposes potential avenues for future research.

2. MATERIALS AND METHODS

Our experimental methods utilize deep learning models (i.e., as Variational Autoencoders) designed to encode timbral information from audio into a latent space. Additionally, we introduce a novel timbre dataset specifically created to capture the subtle timbral variations produced by diverse extended tenor saxophone techniques.

2.1 RAVE models

Realtime Audio Variational AutoEncoder (RAVE) models are deep learning models specifically crafted for generating high-quality audio in real-time [4]. Unlike traditional Variational Autoencoders (VAEs), which can struggle with reconstruction accuracy, RAVE uses a twofold training process (representation learning and adversarial fine-tuning) for enhanced performance. It breaks down audio into multiple frequency bands, capturing fine degrees of detail from large ($\approx$ 48kHz) sampling rate input audio representations without sacrificing efficiency. RAVE models can be used in timbre transfer (i.e., change the timbre of a sound from, e.g., a piano into a violin), for manipulating audio in real-time for performances, or even for compressing audio files.

We begin our exploration with pre-trained RAVE models, detailed in Table 1, originating from both the Institut de Recherche et Coordination Acoustique/Musique (IRCAM) and the Intelligent Instruments Lab (IIL). These models are trained on various raw waveform datasets, as summarized in Table 1, providing a foundation for investigating their capacity to generalize across diverse musical timbres.

The models are built upon different RAVE architectures: IRCAM's models are based on RAVE v2, an enhanced continuous model compared to the one outlined in [4], employing the Variational Auto Encoder objective (ELBO or evidence lower bound function) for regularization. Conversely, IIL's models are trained on a modified iteration of RAVE v1, facilitating transfer learning capabilities. The IIL models exclusively consist of encoder-decoder configurations, without priors, utilizing causal convolutions and optimized for streaming inference with MAX/MSP, Puredata (PD), and Supercollider [18].

Notably, IRCAM's models operate at a sampling rate of 44.1kHz, while IIL's models operate at 48kHz. Furthermore, IIL's models typically feature a greater number of layers, averaging between 16 to 20 layers compared to IRCAM's 8 to 16 layers.

2.2 Tenor Saxophone Timbre Dataset

In this study, our aim was to delve into a timbre latent space, enabling us to capture the nuanced tonal intricacies emanating from the sonic palette of the tenor saxophone. Renowned for its versatility within the timbral spectrum, this instrument offers a plethora of well-documented extended techniques, as documented by Kientzy [25] and Weiss [26]), extensively explored in the context of contemporary Western Art Music.

To document the timbre of the tenor saxophone, we have curated a collection of instrumental tones encompassing a broad array of combinations of (extended) techniques, registers, and dynamics. The details of these combinations are summarized in Table 2. We recorded 191 samples using a stereo pair of microphones (Sontronics STC-1S MT BK) at a sample rate of 48 kHz. The sounds are sampled at 44.1 kHz with a resolution of 24 bits. We provide a more comprehensive list, as well as the dataset, as supplementary material in `https://acesse.one/tenor-saxophone-dataset`.

Extended techniques entail the utilization of the instrument in unconventional manners by expanding the sonic palette [27]. The final set of conventional and extended techniques adopted in our dataset are the following: (1) long non-vibrato notes, and (2) long vibrato notes; (3) long quarter-tone notes; (4) long subtone notes, characterized as an "airy, breathy way of playing lower notes" [26]; (5) flutter-tonguing notes – described as "a kind of tremolo on one tone (...) created with the tip of the tongue" [26]; (6) slap notes – described as "an especially sharp tonguing of a percussive character" [26]; (7) bisbigliando – interpreted as a color trill; (8) multiphonics, resulting from two or more simultaneous pitches sounding on what is otherwise a monophonic instrument; and (9) growling long notes – simultaneous playing and singing.

Due to the unique characteristics of each technique, we accounted for variations in register and dynamics. For instance, in the case of long notes, both with and without vibrato, and quarter-tone long notes, we concentrated on the four different pitch registers previously described – Altissimo, High, Medium, and Low – including two notes for each of these registers: C7 and G6 (altissimo), F6 and C6,(high), D5 and A4 (medium), and C4 and Bb3 (low). [1]

[1] These notations are specifically tailored for tenor saxophone tuned to

Name	Team	RAVE	Sample Rate	Lat. Dim.	Dataset	Ref.
Sol_full	IRCAM	v2	44.1kHz	8	Full Studio OnLine IRCAM database, incorporating over 20,000 samples of extended instrumental playing techniques	[19, 20]
Isis	IRCAM	v2	44.1kHz	8	ISiS Database, comprising synthesized high-quality singing voices, akin to real singers	[21]
MusicNet	IRCAM	v2	44.1kHz	16	Musicnet database, comprising recordings spanning 34 hours of chamber music performances, showcasing 11 instruments	[22, 23]
IILGuitarTimbre	IIL	v1	48kHz	16	Timbre-oriented collection of plucking, strumming, striking, scraping, and more recorded dry from an electric guitar	[18]
SaxSoprano	IIL	v1	48kHz	20	Original soprano saxophone improvisation by Franziska Schroeder	[18]
VocalSet	IIL	v1	48kHz	16	VocalSet dataset, comprehending high-quality recordings of 20 professional singers exhibiting diverse vocal techniques across various vowels	[18, 24]

Table 1. Pre-trained RAVE models employed in our experiments, their characteristics (version, sample rate, and number of latent dimensions), and dataset description.

Technique	Register	Dynamic
Long non-vibrato notes	Altissimo	Very Loud
Long vibrato note	High	Loud
Quarter-Tones	Medium	Medium
Flutter-tonguing	Low	Soft
Slap notes	Very Soft	Very Soft
Growling		
Bisbigliando		
Subtones		
Multiphonics		

Table 2. Collection of sounds encompassing the entire spectrum of the tenor saxophone (timbre space).

Simultaneously, each note was performed at different dynamic levels – Very Loud, Loud, Medium, Soft, and Very Soft.

Given the specific characteristics of the technique, long subtone notes are exclusively applied to the low register of the instrument, specifically notes Low D4, Low C4, Low B3, and Low Bb3. Due to their challenging execution and infrequent usage, flutter-tonguing notes are omitted from the altissimo register (above G6). Additionally, the two notes of each High, Medium, and Low register are studied without dynamic variation.

Efficiency constraints similarly result in the omission of slap notes from the altissimo register. Various registers (High, Medium, Low) were incorporated, each with two notes, featuring variations between a drier attack (secco slap) and a standard attack with clear pitch recognition. Regarding bisbigliando, we established position variations for the established notes for each register (Altissimo, High, Medium, and Low).

In the case of multiphonics, we recorded the following five multiphonics [26, 28] addressing different notes and registers without introducing dynamic variations:

1. C#5, D6 (quarter-tone higher), A6 [2]
2. D5 (three quarter-tone higher), F#5 [3]
3. F#4, G5 [4]
4. D4 (quarter-tone lower), Bb5, D6 (quarter-tone higher), F#6 [5]
5. D5 (quarter-tone higher), G5 (quarter-tone higher), E6, B6 [6]

For Growling Long Notes, we gathered notes distributed across various registers (Altissimo, High, Medium, and Low) without dynamic variations. This decision was made due to the inherent performance limitations of this technique, particularly in soft or very soft dynamics.

These sounds represent a diverse range of timbral qualities and extended techniques inherent to the tenor saxophone, offering potential avenues for exploring the timbre latent space.

3. EVALUATION METHODOLOGY

To explore the transparency afforded by navigating the latent space created by RAVE models within the timbre

B-flat. In concert pitch, the corresponding notes are altissimo (Bb5 and F5), high (Eb5 and Bb4), medium (C4 and G3), and low (Bb2 and Ab2).

[2] concert pitch: B3, C5 (quarter-tone higher), G5
[3] concert pitch: C4 (three quarter-tone higher), E4
[4] concert pitch: E3, F4
[5] concert pitch: C3 (quarter-tone lower), Ab4, C5 (quarter-tone higher), E5
[6] concert pitch: C4 (quarter-tone higher), F4 (quarter-tone higher), D5, A5

domain of the tenor saxophone, we adopt a two-pronged strategy. Firstly, a subjective perceptual assessment aims to offer a comprehensive understanding of the sensitivity of RAVE model layers to various sonic attributes within the context of the tenor saxophone's timbre characteristics. Secondly, an objective evaluation utilizes clustering metrics to analyze the degree to which layers demonstrate sensitivity to multiple sound attributes. This involves assessing whether layers tend to group specific attributes, indicating a heightened sensitivity towards them.

3.1 Perceptual Subjective Assessment

The first evaluation approach delves into the complex layers of the mentioned RAVE model's encoder and decoder outputs, aiming to serve as a guide for systematic empirical analysis.

The perceptual subjective assessment was conducted by the saxophonist Jorge Sousa, who has thirteen years of professional experience playing the tenor saxophone, seeking to provide an in-depth understanding of the sensitivity of RAVE model layers to multiple sonic attributes within the context of the timbre characteristics of the tenor saxophone. This evaluation strives to unveil the unique transformations and representations at various stages of the encoding and decoding processes.

In this scenario, every sound within the dataset undergoes meticulous examination, with thorough listening sessions conducted at each individual layer of every model enumerated in Table 1. The output of each neuron in every layer was converted into a sound signal and listened to individually. The analysis was realized in the visual multimedia programming language Pure Data (PD), [7] using IRCAM's nn_tilde object for PD. [8]

3.2 Analysis of Timbre Clusters

The second evaluation approach objectively assesses the extent to which the models' latent spaces recognize semantically meaningful musical attributes (i.e., timbre, pitch, and dynamics). This assessment quantifies intra- and inter-segment distances across various parameters, such as technique, register, dynamic, and their combination, across all sounds in the provided dataset.

To this end, we construct the latent space for individual sounds, sampled in segments with 40 milliseconds of duration. Clustering is then partitioned into four primary categories: (1) technique, (2) pitch register, (3) dynamic, and (4) all combined. For the initial two categories, we employ metrics on the sounds both with and without normalization, aiming to ascertain whether the overall inherent dynamics of each technique and register significantly influence the model's capacity to discern the attributes under investigation. For normalization, we use librosa [9] to scale the audio input to have a maximum absolute value of 1, ensuring the signal fits within a standard dynamic range.

We utilize two cluster evaluation metrics, the Davis-Bouldin score (DB) [29] and Dunn index [30], to assess both intra-segment distances within each cluster and inter-cluster distances across the latent spaces. The Davis-Bouldin score and Dunn index evaluate the resulting latent space's ability to produce compact and well-separated sound clusters. A lower Davis-Bouldin score indicates better-separated clusters, while a higher score indicates poorly separated clusters. The Dunn index aims for maximization, with the intra-cluster metric (cluster diameter) calculated as the average Euclidean distance across all cluster segments and the inter-cluster metric determined as the distance between each cluster's nearest neighbors. A higher Dunn index indicates more distinct and tightly packed clusters. In this context, the metric compares which layer better discriminates a certain attribute without considering minimum or maximum expected values.

To enable a more objective analysis across the output of each layer, we implement this method for the entire latent space of each sound within the clusters and every individual layer, supplemented by the findings from perceptual assessments. This thorough evaluation aims to uncover the underlying mechanisms governing the representation of tenor saxophone timbre within the RAVE framework. We make the code available at `https://github.com/NadiaCarvalho/SMC-TimbreLandscape.git`.

4. RESULTS AND DISCUSSION

4.1 Perceptual Subjective Assessment

Observations drawn from subjective perceptual assessments of each layer's distinct responses to the sound samples of the tenor saxophone offer valuable perspectives that we can integrate into our objective evaluation process.

For Model **Sol_full**, Jorge's examination indicates that six of the eight layers exhibit no discernible reaction to the instrument's sound samples: layers two, four, five, six, seven, and eight. Layers one and three consistently show minimal differentiation among different pitches and dynamics.

The **Isis** model displays notable reactivity to louder dynamics (Loud and Very Loud) globally, although it is less efficient in responding to various multiphonics. It shows heightened responsiveness in the lower register of the tenor saxophone, particularly with notes such as Bb3, B3, and C4. [10]

Regarding the model trained on **Musicnet**, layer one demonstrates reactivity to nearly all tenor saxophone sound samples, particularly excelling in discerning and responding to samples with distinctive attacks (articulation in slap). However, layers fourteen, fifteen, and sixteen, especially the last one, prove ineffective at discerning most sounds. Layer five stands out for its ability to modify the pitches of introduced sounds.

Model **IILGuitarTimbre** shows layer one's reactivity to all tenor saxophone sound samples, albeit with minimal

[7] `https://msp.ucsd.edu/software.html`, accessed on March 10, 2024.

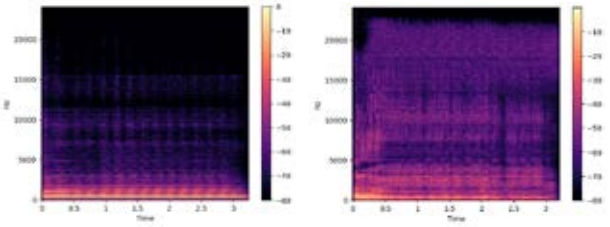
[8] `https://github.com/acids-ircam/nn_tilde`, accessed on March 10, 2024.

[9] `https://librosa.org`, accessed on March 10, 2024.

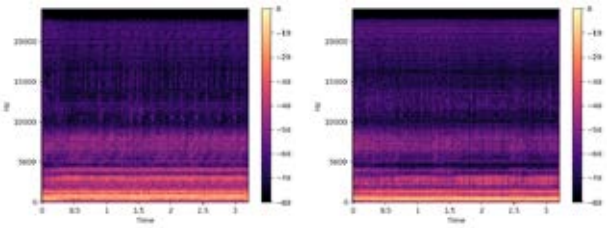
[10] Pitch in this section is always presented in the notation specifically tailored for tenor saxophone tuned to B-flat.

responsiveness to changes in dynamics and pitch. Conversely, layers two, five, ten, thirteen, and fourteen show no response to any stimulus. Layer seven demonstrates noticeable reactions to sound stimuli at low dynamics (Soft and Very Soft).

The **SaxSoprano** model is generally deemed ineffective. Among its twenty layers, eleven show no reaction to tenor saxophone sound samples. Layer one emerges as the most responsive, consistently reacting across various dynamics and effectively responding to slap articulations and multiphonics. Layer eleven consistently reacts to the note D5 (medium register), a unique trait not found in other models. Activating layer one influences other layers, leading to distinct behaviors, such as amplifying low frequencies, adding granular texture, producing a "Pan flute" sound, and closely resembling the input pitch.



(a) Input Audio (b) Layer 1

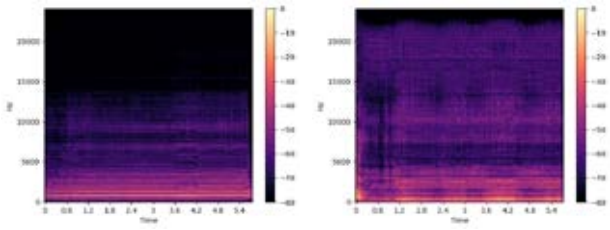


(c) Layer 11 (d) Layer 16

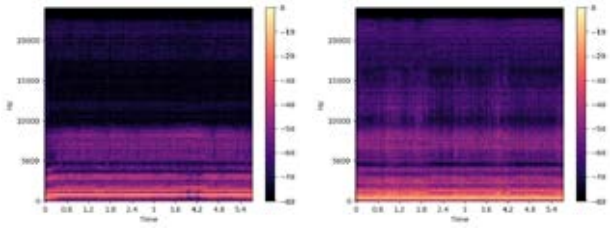
Figure 1. Spectrograms for the input audio of the soft low C recording, as well as the sounds decoded from layers 1, 11, and 16 of the latent space.

In our examination of the **VocalSet** model comprising sixteen layers, we found it to be a balanced model, with layers responding to sound samples across all registers (from Low to Altissimo), except for layer sixteen, which showed greater reactivity to the Low register (as observed in Fig. 1). While several layers reacted to multiphonics (see Fig. 2), the second and third multiphonics didn't elicit any response. Layer one emerged as the most reactive layer, layer eleven displayed high reactivity to the Altissimo register, and layer thirteen responded solely to note C across various registers.

Finally, among the models studied, we propose that the Musicnet and Vocalset models are the most reactive to tenor saxophone sound samples. Despite both being saxophones, the model trained on soprano saxophone sounds proves ineffective when examined layer-by-layer. Layer one consistently exhibits the most robust reaction to various registers, pitches, effects, and dynamics. The models demonstrate higher reactivity to louder dynamics (Loud or Very Loud) and are generally more effective in lower



(a) Input Audio (b) Layer 1



(c) Layer 10 (d) Layer 16

Figure 2. Spectrograms for the input audio of the multiphonic number five, as well as the sounds decoded from layers 1, 11, and 16 of the latent space.

notes, particularly Bb3. Additionally, we observe an apparent inability to react to and distinguish between various presented multiphonics.

4.2 Analysis of Timbre Clusters

Table 3 shows the cluster analysis results, encompassing all layers of each model. From these findings, several conclusions emerge. Foremost among them is the clear indication that the absence of normalization negatively impacts the Davis-Bouldin score when applying clustering metrics to extended techniques and pitch registers across all models. Consequently, the models struggle to effectively discern extended techniques, and consequently, timbre changes or register variations when dynamics vary. The exception to this trend is observed in the VocalSet model, where both normalized and non-normalized audio inputs yield very similar values for both technique and pitch register.

Among the models evaluated, the SaxSoprano model stands out for its superior ability to distinguish between extended techniques, particularly when the audio is normalized. With the second-best Davis-Bouldin score and the highest Dunn index score, it demonstrates notable proficiency in this aspect. A noteworthy finding concerns the MusicNet model, which achieves the best Davis-Bouldin score but one of the lowest Dunn index scores. This suggests that while the MusicNet model excels in forming distinct timbre clusters due to its capacity to capture unique data patterns, it struggles to achieve adequate separation between these clusters. Consequently, it faces challenges differentiating characteristics to create clearly distinct timbre clusters.

Regarding pitch registers, most of the models perform similarly with normalized audio. The exception is the IILGuitarTimbre, which presents very high values of the Davis-Bouldin score, meaning it cannot distinguish pitch

Model		Technique		Register		Dynamic	All
		N	NN	N	NN		
Sol_full	**DB**	9,3978	63,1686	3,1654	71,0053	61,2390	23,3833
	Dunn	0,0001	0,0003	0,0002	0,0003	0,0002	0,0014
Isis	**DB**	7,8626	36,9219	**2,5262**	17,9766	30,7508	18,4759
	Dunn	0,00002	0,0003	0,0001	**0,0005**	0,0001	0,0006
MusicNet	**DB**	**6,9871**	27,2322	3,4207	22,0426	35,3139	14,3542
	Dunn	0,00002	0,0003	0,0001	**0,0005**	**0,0003**	0,0016
IILGuitarTimbre	**DB**	46,4832	69,0797	25,9224	40,3061	58,2117	35,3975
	Dunn	0,0002	0,0002	**0,0005**	0,0003	0,0001	0,0003
SaxSoprano	**DB**	8,8365	23,6323	3,1240	5,8060	14,4397	9,2430
	Dunn	**0,0003**	**0,0004**	0,0004	**0,0005**	0,0002	**0,0034**
VocalSet	**DB**	11,4939	**12,7429**	2,5283	**2,9858**	**8,4275**	**5,5918**
	Dunn	**0,0003**	0,0003	0,0002	**0,0005**	**0,0003**	0,0024

Table 3. Clustering Results Across All Layers. The best results for each clustering category and metric are marked in bold. N and NN denote normalized and non-normalized audio, respectively.

well. However, contrary to the previous example regarding extended techniques, it features the best Dunn index. This suggests that the dataset used for training may encompass greater complexity regarding pitch registers associated with extended techniques. This distinction might be attributed to the fact that the IILGuitarTimbre model is trained on a dataset featuring sounds more distant from those of the tenor saxophone, unlike the other models, which are associated with instrumental orchestral and chamber music sounds or voice, often resembling saxophone timbres.

Concerning dynamics (coupled with extended techniques) and all sounds clustered separately, the VocalSet model emerges as the top performer for both metrics, closely followed by the SaxSoprano model.

Figure 3 showcases the outcomes of the cluster analysis, organized layer by layer. Due to the close resemblance of Dunn index results across all metrics (approximately 1E-05), only Davis-Bouldin score results are presented in the figure. Several discernible patterns emerge from these results, complementing the listening test findings. Specifically, dynamics are consistently better generalized in the initial layer of the model, except in the case of the IILGuitarTimbre and VocalSet models, where they are observed in the second and sixth layers, respectively. Combined sounds are not adequately clustered at any model layer, likely due to the limitation of a single layer in capturing the diverse timbre attributes of entire sounds. The pitch register is generally well perceived across the models, although not consistently at the same layer in every model. At the level of extended techniques, Musicnet, closely followed by VocalSet and Isis, are the best models at semantically perceiving timbric changes by a single layer.

5. CONCLUSIONS AND FUTURE WORK

This study conducts a comprehensive examination of RAVE models to uncover the nuances of their latent spaces. Emphasizing a specialized timbre latent space tailored for saxophone extended techniques, we delve into the encoder and decoder outputs of RAVE layers. Through a dual evaluation method, we subjectively analyze outputs layer by layer and compare them across diverse model datasets. Additionally, we objectively assess the models' encoding latent spaces' ability to recognize semantically meaningful timbre attributes. This assessment involves quantifying intra- and inter-segment distances across various sound attributes, including technique, register, and dynamics, across the entire dataset.

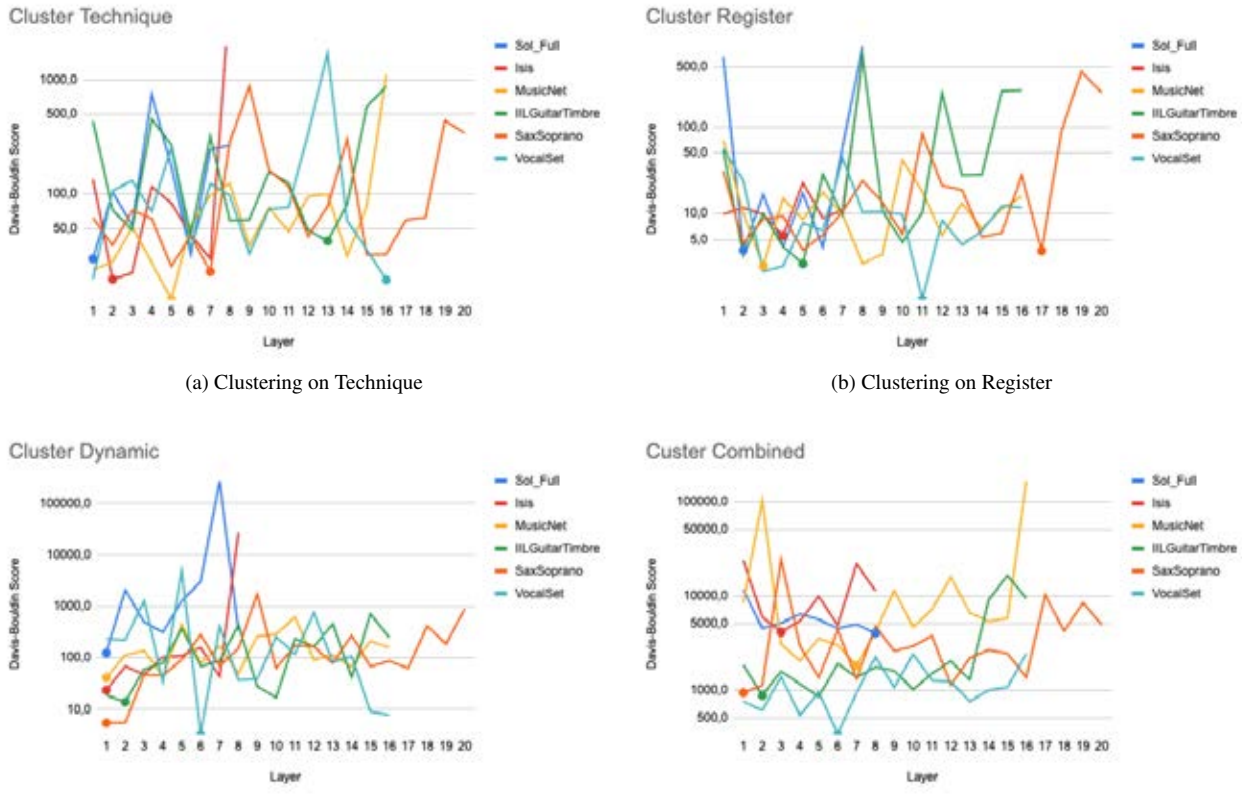
The conclusions drawn from this investigation into RAVE models reveal crucial insights into their ability to capture and represent saxophone timbre attributes. Through a meticulous dual evaluation approach encompassing subjective listening tests and objective clustering analyses, several key findings emerge.

Firstly, among the models studied, MusicNet and VocalSet exhibit heightened reactivity to tenor saxophone samples, particularly evident in the subjective perceptual layer-by-layer examination. However, the effectiveness varies across models, with the soprano saxophone model proving less reactive. Notably, the initial layer consistently demonstrates robust responses to different attributes, highlighting its pivotal role in timbre representation.

Secondly, the absence of normalization negatively impacts clustering metrics, hindering the models' ability to effectively distinguish extended techniques and pitch registers. While some models, such as SaxSoprano, excel in distinguishing between extended techniques, others struggle to achieve adequate separation, as seen in the case of MusicNet. Furthermore, dynamics and combined sounds present additional challenges in clustering, underscoring the complexity of timbre representation.

Lastly, objective layer-by-layer analysis complements the findings from perceptual evaluation, uncovering consistent patterns in dynamics perception across models. While there are variations in pitch register perception across different layers, MusicNet and VocalSet stand out as top performers in semantically perceiving timbral changes within a single layer.

In summary, the comprehensive evaluation elucidates fundamental mechanisms governing tenor saxophone timbre representation in RAVE models. The insights gleaned



(a) Clustering on Technique

(b) Clustering on Register

(c) Clustering on Dynamic

(d) Combined Clustering

Figure 3. Davis-Bouldin Score Results by Layer. The best results are marked with a circle. All audio is normalized for the cluster categories corresponding to extended techniques and pitch registers.

contribute to advancing our understanding of neural network-based timbre modeling and hold implications for music synthesis and analysis applications.

Moving forward, we plan to further investigate the RAVE model layers by analyzing their behavior across individual registers, dynamics, and extended techniques. Such an endeavor promises to unveil further insights into the intricacies of RAVE models, thereby enhancing transparency in their functioning. Additionally, we aspire to move beyond mere transparency and innovate new tools that empower users to harness the potential of these models creatively. By enabling users to take control and experiment with the models, we envision facilitating the creation of music and performances that are enriched and elevated. This forward-looking approach not only advances our understanding of RAVE models but also fosters a dynamic and collaborative environment for music exploration and expression.

Acknowledgments

This research has been funded by the Portuguese National Funding Agency for Science, Research and Technology [2021.05132.BD and 2023.01345.BD].

6. REFERENCES

[1] H. Scurto and A. Chemla-Romeu-Santos, "Deeply Listening Through/Out the Deepscape," in *28th International Symposium on Electronic Art (ISEA 2023)*, Paris, France, May 2023. [Online]. Available: https://hal.science/hal-04108995

[2] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, and R. B. Grosse, "Timbretron: A wavenet(cyclegan(cqt(audio))) pipeline for musical timbre transfer," *Computing Research Repository (CoRR)*, vol. abs/1811.09620, 2018. [Online]. Available: http://arxiv.org/abs/1811.09620

[3] D. K. Jain, A. Kumar, L. Cai, S. Singhal, and V. Kumar, "Att: Attention-based timbre transfer," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–6.

[4] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *Computing Research Repository (CoRR)*, vol. abs/2111.05011, 2021. [Online]. Available: https://arxiv.org/abs/2111.05011

[5] 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 *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1068–1077. [Online]. Available: https://proceedings.mlr.press/v70/engel17a.html

[6] F. Roche, T. Hueber, S. Limier, and L. Girin, "Autoencoders for music sound modeling : a comparison of linear, shallow, deep, recurrent and variational models," in *SMC 2019 - 16th Sound & Music Computing Conference*, ser. Proc. of SMC 2019, U. of Malaga (UMA), Ed., no. 1-6, Malaga, Spain, May 2019. [Online]. Available: https://hal.science/hal-02349406

[7] K. Tatar, D. Bisig, and P. Pasquier, "Latent timbre synthesis: Audio-based variational auto-encoders for music composition and sound design applications," *Neural Computing and Applications*, vol. 33, no. 1, p. 67–84, Oct. 2020. [Online]. Available: http://dx.doi.org/10.1007/s00521-020-05424-2

[8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *Computing Research Repository (CoRR)*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[9] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. C. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *Computing Research Repository (CoRR)*, vol. abs/1612.07837, 2016. [Online]. Available: http://arxiv.org/abs/1612.07837

[10] K. Kumar, R. Kumar, T. de Boissière, L. Gestin, W. Z. Teoh, J. M. R. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Neural Information Processing Systems*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1910.06711

[11] S. Nercessian, "P-rave: Improving rave through pitch conditioning and more with application to singing voice conversion," in *Proceedings of the 26th International Conference on Digital Audio Effects (DAFx23)*, 2023.

[12] F. Schroeder and F. Reuben, "Ai Improvised Music Duo," *AIMC 2023*, aug 29 2023, https://aimc2023.pubpub.org/pub/8x9jxz9a.

[13] J. C. Reus, "i o we," *AIMC 2023*, aug 29 2023, https://aimc2023.pubpub.org/pub/hpy32yre.

[14] J. Armitage, "Gagnavera," *AIMC 2023*, aug 29 2023, https://aimc2023.pubpub.org/pub/ppgwfht1.

[15] ——, "Strengjavera," *AIMC 2023*, aug 29 2023, https://aimc2023.pubpub.org/pub/83k6upv8.

[16] A. Chemla–Romeu-Santos and P. Esling, "Challenges in creative generative models for music: a divergence maximization perspective," in *Proceedings of the 3rd Conference on AI Music Creativity*. AIMC, Sep. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7088272

[17] T. Letowski, "Timbre, tone color, and sound quality: concepts and definitions," *Archives of Acoustics*, vol. 17, pp. 17–30, 1992.

[18] Intelligent Instruments Lab, "rave-models (revision ad15daf)," 2023. [Online]. Available: https://huggingface.co/Intelligent-Instruments-Lab/rave-models

[19] G. Ballet, R. Borghesi, P. Hoffmann, and F. Lévy, "Studio Online 3.0: An Internet "Killer Application" for Remote Access to IRCAM Sounds and Processing tools," in *Journées d'Informatique Musicale*, Paris, France, May 1999. [Online]. Available: https://hal.science/hal-03112091

[20] C. Cella, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg, and Y. Maresz, "Orchideasol: a dataset of extended instrumental techniques for computer-aided orchestration," *Computing Research Repository (CoRR)*, vol. abs/2007.00763, 2020. [Online]. Available: https://arxiv.org/abs/2007.00763

[21] L. Ardaillon, "Synthesis and expressive transformation of singing voice," Theses, Université Pierre et Marie Curie - Paris VI, Nov. 2017. [Online]. Available: https://hal.science/tel-01710926

[22] J. Thickstun, Z. Harchaoui, and S. Kakade, "Learning features of music from scratch," 2017.

[23] J. Thickstun, Z. Harchaoui, and S. M. Kakade, "Musicnet," Jul. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5120004

[24] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. H. 0001, E. Humphrey, and E. Benetos, Eds., 2018, pp. 468–474. [Online]. Available: http://ismir2018.ircam.fr/doc/pdfs/114_Paper.pdf

[25] D. Kientzy and D. Charles, "Saxologie du potentiel acoustico - expressif des 7 saxophones," Ph.D. dissertation, Paris 8, 1990.

[26] M. Weiss and G. Netti, *The Techniques of Saxophone Playing*. Bärenreiter, 2010.

[27] M. Burtner, "Making noise: Extended techniques after experimentalism," *New Music Box*, vol. 6, 2005.

[28] D. Kientzy, *Les Sons Multiples aux Saxophones*. Editions Salabert, 1982.

[29] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[30] J. C. Dunn†, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974. [Online]. Available: https://doi.org/10.1080/01969727408546059