DMRN+15: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2020

QUEEN MARY UNIVERSITY OF LONDON
TUE 15 DECEMBER 2020

Perceptual Similarities in Neural Timbre Embeddings

Ben Hayes, Luke Brosnahan, Charalampos Saitis, and George Fazekas*

Centre for Digital Music
Queen Mary University of London, United Kingdom
b.j.hayes@qmul.ac.uk

***Abstract*— Many neural audio synthesis models learn a representational space which can be used for control or exploration of the sounds generated. It is unclear what relationship exists between this space and human perception of these sounds. In this work, we compute configurational similarity metrics between an embedding space learned by a neural audio synthesis model and conventional perceptual and semantic timbre spaces. These spaces are computed using abstract synthesised sounds. We find significant similarities between these spaces, suggesting a shared organisational influence.**

***Index Terms*—** Neural audio synthesis, psychoacoustics, timbre, representation learning

I. INTRODUCTION

Many neural audio synthesis models use representation learning techniques to enable interpretable control. For example, Kim *et al* learned an instrument embedding when training their *Mel2Mel* model, in a manner that required only reconstruction loss [1]. In this work, we compare the organisation of a *Mel2Mel* embedding space with perceptual and semantic timbre spaces computed from human ratings.

II. METHOD

We use a set of twelve sounds created with frequency modulation (FM) synthesis in a previous study [2]. Participants ($n=30$) provided pairwise dissimilarity ratings on these stimuli, and an English speaking subset ($n = 24$) provided semantic ratings along 30 adjective scales. Adjectives were sourced by text-mining a corpus from a popular modular synthesis forum. A 3D timbre space was constructed by performing multidimensional scaling (MDS) on the dissimilarity scores. A 2D semantic space was computed with exploratory factor analysis (EFA) on the semantic ratings. Horn's parallel analysis supported two-factors, which was subject to non-orthogonal Oblimin rotation.

Mel2Mel's embedding space is given by a matrix transformation of a one-hot instrument vector. The transformation is learned by backpropogation through two featurewise linear modulation (FiLM) conditioning layers. The organisation of this space is therefore motivated by the network's overall reconstruction objective. Two versions of the model were trained, with 2D and 3D embedding spaces.

*This work was supported by UK Research and Innovation [grant number EP/S022694/1]

Table 1: Configurational Similarity Metrics

Space	Embed.	*T.C.C.*	m^2	*RVmod*
EFA	2D	0.884	0.439[a]	0.683
MDS	3D	0.923	0.721	0.325

[a] PROTEST significance $p<0.001$

III. RESULTS

Three configurational similarity metrics were used. Tucker's congruence coefficient (TCC) is related to the cosine similarity between factors, and is computed after Procrustes rotation. A TCC of $0.83-0.95$ is considered significant, and >0.95 nearly identical [3]. m^2, is the minimisation objective of Procrustes rotation. The modified RV coefficient, is an extension of Pearson's r to matrices. Table 1 shows these metrics for each timbre space and the embedding space of corresponding dimensionality. We see strong similarity across all metrics in the semantic EFA space, and very strong similarity in only TCC in the MDS space.

IV. CONCLUSION

The similarities between the timbre spaces and the *Mel2Mel* embedding spaces suggest that both systems rely on similar attributes to discriminate these sounds. Whilst not conclusive, our results indicate that further work is warranted. This will include investigation into whether these results generalise to other types of sound and other NAS architectures, including those with different representational spaces. The finer structure of these spaces can also be studied by observing the positioning of latent space interpolations in perceptual and semantic timbre spaces.

V. REFERENCES

[1] J. W. Kim, R. Bittner, A. Kumar, and J. P. Bello, "Neural Music Synthesis for Flexible Timbre Control," *arXiv:1811.00223 [cs, eess, stat]*, Nov. 2018.

[2] B. Hayes and C. Saitis, "There's more to timbre than musical instruments: Semantic dimensions of FM sounds," in *Proceedings of the 2nd International Conference on Timbre*, Thessaloniki, Greece (Online), 2020.

[3] U. Lorenzo-Seva and J. M. F. ten Berge, "Tucker's congruence coefficient as a meaningful index of factor similarity," *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, vol. 2, no. 2, pp. 57–64, 2006.