**DMRN+14: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2019**

***QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2019***

Modulation Spectra for Musical Dynamics Perception and Retrieval

Luca Marinelli[1], Athanasios Lykartsis[1], and Charalampos Saitis[2]

[1]Audio Communication Group, TU Berlin, Germany
[2]Centre for Digital Music, Queen Mary University of London, UK, c.saitis@qmul.ac.uk

Abstract— **To investigate variations in timbre space with regard to musical dynamics, a convolutional neural network was trained on modulation power spectra of single notes of sustained instruments played at pianissimo and fortissimo dynamics. Samples were rms-normalized to eliminate loudness information and force the network to focus on timbre attributes of dynamics shared across different instrument families.**

I. INTRODUCTION

Recent research has shown that even if no loudness cues are available, listeners can still quite reliably identify the intended dynamic strength of a performed sound by relying on timbral features [1]. More recently, acoustical analyses across an extensive set of anechoic recordings of instrument notes played at pianissimo (*pp*) and fortissimo (*ff*) showed that attack slope, spectral skewness, and spectral flatness together explained 72% of the variance in dynamic strength across all instruments, and 89% with an instrument-specific model [2]. The overall aim of the research presented here is to further investigate the role of timbre in musical dynamics, focusing on the contribution of spectral and temporal modulations.

II. METHOD

Using 33 sustained instruments from the same database as [2], 1 s snippets were extracted from the steady-state part of notes. The modulation power spectrum (MPS) is implemented as the squared amplitude of the two-dimensional Fourier transform of the logarithmic amplitude of the mel-scaled short time Fourier transform (STFT). For each time frame of the STFT, the rms was computed for the whole frequency range and used to normalize the same frame.

The CNN architecture (Table 1) was implemented with Keras running on top of TensorFlow. Average pooling was chosen because max pooling seemed to promote overfitting. All activation functions, but the softmax on the last dense layer, are rectified linear units. Instead of tuning each model separately, a global setup for all experiments was used.

III. RESULTS

The model obtained an accuracy of 91.7% for brass instruments, 97.3% for single reeds, 85.2% for double reeds, 64.9% for bowed strings, and 92.6% in a 10-fold cross validation for the entire dataset.

Through visualization of the *pp* and *ff* saliency maps of the CNN it was possible to identify discriminant regions of the MPS and define an audio descriptor. A linear discriminant analysis with 10-fold cross validation using this MPS-based descriptor on the entire dataset performed better than using two STFT-based spectral descriptors, namely spectral skewness and spectral flatness (43.2% error reduction).

TABLE I.

CNN Architecture

Layer type	***Parameters***
Conv2D	filters: 16, size: 7x7, stride: 3
Batch norm.	--
Conv2D	filters: 32, size: 3x3, stride: 1
Batch norm.	--
Average pool.	size: 2x2
Conv2D	filters: 64, size: 3x3, stride: 1
Average pool.	size: 2x2
Flatten	--
Dense	neurons: 128
Dropout	p: 0.5
Dense	neurons: 2

Overall, audio descriptors based on different regions of the MPS could serve as sound representation for machine listening applications, as well as to better delineate the acoustic ingredients of different aspects of timbre perception. Future work should expand on impulsive sounds and include different dynamic gradations.

REFERENCES

[1] M. Fabiani and A. Friberg, “Influence of pitch, loudness, and timbre on the perception of instrument dynamics” *J. Acoust. Soc. Am.*, 130(4), 2011, pp. EL193–EL199.

[2] S. Weinzierl, S. Lepa, F. Schultz, E. Detzner, H. von Coler, and G. Behler, “Sound power and timbre as cues for the dynamic strength of orchestral instruments” *J. Acoust. Soc. Am.*, 144(3), 2018, pp. 1347–1355.