MUSICAL DYNAMICS CLASSIFICATION WITH CNN AND MODULATION SPECTRA

Luca Marinelli, Athanasios Lykartsis, Stefan Weinzierl
Audio Communication Group, TU Berlin
luca.marinelli@campus.tu-berlin.de

Charalampos Saitis
C4DM, Queen Mary University of London
c.saitis@qmul.ac.uk

ABSTRACT

To investigate variations in the timbre space with regards to musical dynamics, convolutional neural networks (CNNs) were trained on modulation power spectra (MPS), mel-scaled and ERB-scaled spectrograms of single notes of sustained instruments played at two dynamics extremes (*pp* and *ff*). The samples, from an extensive dataset of several timbre families, were rms normalized in order to eliminate the loudness information and force the network to focus on timbre attributes of musical dynamics that are shared across different instrument families. The proposed CNN architecture obtained competitive results in three classification tasks with all three input representations. In order to compare the different input representations, the test sets in three experiments were partitioned in order to promote or avoid selection bias. When selection bias was avoided, models trained on MPS were outperformed by those trained on time-frequency representations, conversely, those trained on MPS achieved the best results when selection bias was promoted. Low-temporal modulations emerged in class-specific MPS saliency maps as markers of musical dynamics. This led to the implementation of a MPS-based scalar descriptor of timbre that largely outperformed the chosen baseline (44.8% error reduction).

1. INTRODUCTION

Timbre is often described as a complex set of sound features that are not accounted for by pitch, loudness, duration, spatial location, and the acoustic environment [1]. Musical dynamics refers to the perceived or intended loudness of a played note, instructed in music notation as *piano* or *forte* (soft or loud) with different dynamic gradations between and beyond. Previous research has suggested that listeners can recognize musical dynamics even if there are no loudness cues available by relying on timbral features [2, 3]. More recently, Weinzierl et al. [4] extracted audio descriptors of timbre from an extensive set of anechoic recordings of orchestral instrument notes played at pianissimo (*pp*) and fortissimo (*ff*). They found that *attack slope*, *spectral skewness*, and *spectral flatness* [5] together explained 72% of the variance in dynamic strength across all instruments, and 89% with an instrument-specific model.

The overall aim of this study is to further investigate the role of timbre in musical dynamics, focusing on the contribution of spectral and temporal modulations. Specifically, we aimed to explore the use of modulation power spectra in the context of deep representation learning for music information retrieval.

1.1 Convolutional Neural Networks and Audio

Unlike more traditional machine learning systems based on handcrafted features, which require previous knowledge of the studied domain and careful engineering, convolutional neural networks (CNNs) combine feature extraction and classification. Originally used in computer vision, they recently have been shown to be a powerful approach for a variety of music information retrieval tasks.

Using mel-scaled magnitude spectrograms as input, thus interpreting a machine listening problem as a machine vision one, Schlter et al. [6] found that in an elementary task of musical onset detection CNNs outperformed previous state of the art methods while requiring less manual preprocessing. By analyzing their model, they found that it had essentially learned features already implemented in hand-designed descriptors, but it surpassed them by combining the results of many slightly different versions. Han et al. [7] found that the performance of CNNs for the recognition of predominant instruments in polyphonic music was more robust than traditional methods, also using mel-spectrograms as input. Furthermore, Phan et al. [8] showed that a more "shallow" CNN architecture outperformed its deeper counterparts in an audio event recognition task. Their CNN consisted of just three layers: a convolutional layer coupled with a pooling layer for feature extraction, and a final softmax layer for classification. Each convolutional filter can be considered as mimicking a cochlear filter that spikes on a specific feature of the signal.

Inspired by these results, we applied a similar approach to the recognition of musical dynamics. In particular, we were interested whether compact networks like those implemented in this study can provide competitive results without the large power and memory requirements of very deep architectures [8].

1.2 Modulation Power Spectra and Timbre

A less explored two-dimensional representation of sound to be used as input in CNN systems is the modulation power spectrum (MPS). The MPS is obtained through successive Fourier transforms along the time and frequency axes of a spectrogram, which highlight the temporal and spectral

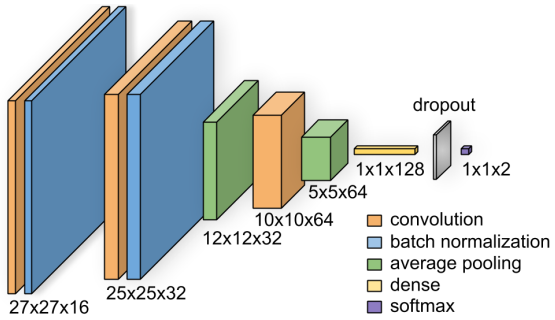

Figure 1: Schematic depiction of the proposed CNN (with feature maps).

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

Table 1: A layer-wise definition of the proposed CNN

regularities of the latter [9]. As such, it offers an invertible representation of the spectrum of a sound that is invariant to translations in the time-frequency domain [10]. Similarly to the spectrogram, the MPS can be associated with a processing stage in the auditory system [11].

Research in psychoacoustics, auditory modelling, and music information retrieval has pointed out the important role of spectrotemporal modulations in the perception and recognition of the timbre of acoustic instrument sounds. Patil et al. [12] showed that modulation representations can achieve remarkably high accuracy in automatic instrument classification, and they also correlate well with perceptual dissimilarity ratings between instrumental timbres. Elliott et al. [13] suggested that these depend on spectrotemporal patterns rather than on purely temporal or spectral features such as attack time and spectral centroid. Using the MPS as acoustic description of timbre dissimilarity dimensions was found to have similar predictive power compared to standard hand-crafted scalar audio descriptors. Thoret et al. [10, 14] showed that instrument timbres of sustained notes can be characterized by specific regions in the MPS, with the most salient cues centred on low temporal and low spectral modulations.

The present study aims to explore whether specific regions in the MPS of sustained instrument sounds are relevant for a timbre-driven classification of musical dynamics when the influence of loudness information is excluded. As a benchmark, we compare the performance of the proposed MPS-CNN model against the standard spectrogram-CNN approach.

2. METHOD

2.1 Dataset

In a recent study on the role of timbre in musical dynamics, Weinzierl et al. [4] acquired under controlled conditions a comprehensive dataset of anechoic recordings of single notes for all instruments of a typical Beethovenian orchestra in their modern and historical form.

The musicians were positioned in the geometrical centre of a spherical 32-microphone array. They were asked to play single notes in *ff* (instruction: "play as loud as possible without sounding unpleasant") and *pp* ("play as soft as possible without allowing the sound to break up") in semitone steps over the entire pitch range required in the standard orchestral repertoire, without vibrato for approximately 3 s per note. From three recordings per pitch and dynamic strength, the most musically convincing version was selected manually. The database, including radiation patterns, sound power measurements, and timbre descriptors is fully available on DepositOnce, the research data repository of TU Berlin [15]. Each sample is further annotated with indices of note onset and offset (all samples), and the beginning and end of the steady-state part (sustained sounds only).

Using 33 sustained instruments from the same database, we extracted 1 s snippets starting at the annotated onset of the steady-state part of the corresponding *ff* and *pp* notes. Although a majority of the recorded samples is longer than one second, one-second-excerpts seemed to offer a proper trade-off between input size and classification performance. For further processing, only one of the 32 channels was used from each recording. Calculating a sum of the channels was not considered to avoid comb filter effects. Instead, similarly to Weinzierl et al. [4], for each instrument we selected that channel which most often exhibited the highest rms signal level over all recorded notes as the principal channel (i.e., as the principal direction of sound radiation).

Our dataset thus consists of 2793 mono audio recordings (1362 *ff* and 1431 *pp* notes), originally recorded at a sampling frequency of 44.1 kHz then downsampled to 22.05 kHz. The difference between the size of the *ff* and *pp* note datasets (2.5% of the total dataset) is attributable to the fact that, for some of the samples in the dataset, the length of the sustained phase was shorter than one second.

2.2 Audio Preprocessing

Modulation Power Spectra

The MPS can be defined as the two-dimensional Fourier transform of a spectrogram. As proposed in [14], first the STFT was computed with a hamming window of length 1024 and a hop size of 256. Then, to eliminate the loudness information in further processing, each time frame of the STFT was normalized by its rms which was, therefore, computed along the frequency axis. Subsequently, the frequency axis of the STFT was compressed into the mel-scale

Test sets	mel-spec	ERB-spec	MPS
Individual instruments			
Clarinet	**0.99**	0.97	0.95
Flute	0.60	**0.69**	0.53
Oboe	0.83	**0.87**	0.81
Trombone	**0.99**	0.89	0.81
Violin	0.94	**1.00**	0.90
Instrument families			
Brass	0.88	0.83	**0.91**
Double reeds	0.80	0.76	**0.86**
Single reeds	0.97	0.96	**0.98**
Violin family	0.58	0.55	**0.65**

Table 2: F1 scores (micro avg.) of the CNN models on the individual instrument and instrument family test sets.

using 174 bands, which reduced its size to 174x87.

Finally, the MPS is here implemented as the squared magnitude of the two-dimensional Fourier transform of the logarithmic magnitude of the mel-scaled STFT. Given that the two-dimensional Fourier transform is inherently conjugate symmetric, only the upper half of the resulting MPS was conserved. The resulting two dimensions of the input fed into the CNN (with a size of 87x87) are *temporal modulations* (in Hz; abscissa) and *spectral modulations* (in cycles/Hz; ordinate).

Mel-Scaled Spectrograms

Using the librosa python library [16], a reduction of the size of a magnitude spectrogram (STFT) is performed, where to combine FFT bins into mel-frequency bins a matrix product is computed between a weighting 'filterbank' matrix and the spectrogram, $mel\_filterbank \cdot STFT$, for a predefined number of overlapping approximated filters with a logarithmic increase in bandwidth. To maintain the same size, the mel-scaled spectrograms are again computed with the same parameters and steps (including the rms normalization of the STFT) but with 87 bands this time, resulting again in a input size of 87x87.

ERB-Scaled Spectrograms

With a python implementation [17] of Slaney's Auditory Toolbox [18], a bank of 87 gammatone filters with increasing bandwidth and center frequency ($f_{min} = 20Hz$) is constructed and applied to the raw audio signal to obtain a spectral representation similar to the mel-scaled STFT. Same frame size and hop size were used as for the previous audio representations. The resulting spectrogram was then normalized frame-wise by its rms. The resulting size is, again, 87x87.

2.3 CNN Architecture

The proposed CNN architecture (Fig. 1 and Table 1) was implemented with Keras running on top of TensorFlow. Given that no previous work on MPS as input for CNNs was available, the proposed architecture was selected through systematic trial and error, where the selection criterion was the validation accuracy. Discarded configurations, such as deeper networks, promoted overfitting. The use of strided convolution in the first layer reduced drastically the number of learnable parameters, while in subsequent layers average pooling offered a better performance than strides bigger than 1. Average pooling, with a size of 2x2, was chosen because max pooling seemed to promote overfitting. The proposed architecture includes three 2D-convolutional layers: the first with 16 filters, a 7x7 receptive field and a stride of 3, the second and third with 32 and 64 filters respectively, with a 3x3 receptive field and a stride of 1. Finally, the softmax classifier receives its inputs from a dense layer with 128 neurons. All activation functions, but the softmax on the last dense layer, are rectified linear units.

Input	F1 score
mel-spec	$\mathbf{0.95 \pm 0.007}$
ERB-spec	0.93 ± 0.025
MPS	0.93 ± 0.014
Models	**Wilcoxon p**
MPS Vs mel-spec	0.025
MPS Vs ERB-spec	0.510
mel- Vs ERB-spec	0.005

Table 3: F1 scores (mean $\pm$ standard deviation) and Wilcoxon p values of the CNN models on the entire cross validated dataset.

The training was performed by minimizing the categorical cross-entropy cost function with the Adam-optimizer (with an initial learning rate of 7×10^{-4}). All models were regularized with a dropout layer ($p = 0.5$) after the first dense layer. Two callback functions were used to prevent overfitting, the early stopping method, with a patience of 5 epochs, and the model checkpoint, to save the best model for each training session; both of them monitored the validation loss, where the validation sets were a randomly selected 10% of the training sets. For the sake of comparability, instead of tuning each model separately, a global set-up for all experiments was used.

3. EXPERIMENTS AND RESULTS

3.1 Individual Instruments

In a first round of experiments, *pp* and *ff* notes of the following representative instruments from different timbre families are chosen to define five test sets: clarinet, flute, oboe, tenor trombone and violin. The clarinet, tenor trombone, and violin test sets comprise the historical and modern version of the instrument, while the oboe test set includes the Romantic historical oboe as well. The flute test set contains the Baroque historical traverse flute, historical keyed flute, and the modern transverse flute. For each individual test set a different model was trained. Training and test sets contain mutually exclusive data, and while all *aerophones* samples of the dataset (flute set) are contained in the test set, for all other instruments, the test sets are representative of the respective training sets: for example the model tested on the violin was trained also on samples of cellos, double basses and violas. The class distribution of each test set (*pp*

and *ff*) was the following: clarinet (85, 85), flute (60, 59), oboe (87, 95), tenor trombone (88, 81) and violin (81, 81).

Table 2 (upper panel) reports similarly promising results for all three chosen audio representations. High accuracies were obtained with all representations on clarinet, violin and trombone test sets, where the highest results were achieved with mel-spectrograms or ERB-spectrograms. As expected, the worst performance was obtained on the flute set, given that the training set for the flute did not contain any samples of that timbre family (aerophones instruments) and apparently, only with ERB-spectrograms the CNN was able to extract robust features for the two dynamics extremes.

3.2 Instrument Families

In this second series of experiments, again, mutually exclusive data were used to build training and test sets. This time, similarly to what happened for the aerophones in the previous experiments, the four test sets consisted of all instruments from a specific timbre family. Dividing training and test data in such way aims at assessing which audio representation would provide the most robust features in an experiment that promotes *selection bias*, i.e., where the training sets are ensured to be non-representative of the test sets.

The CNN was tested on four sets comprising *pp* and *ff* notes from both modern and historical versions of instruments belonging to one of the following timbre families: brass, double reeds, single reeds, and bowed strings. In other words, this experiment was conducted as a four-fold cross-validation, where each fold comprised all samples of a certain timbre family, thus strongly introducing selection bias for that particular family. The brass set included two versions of trumpet, French horn, tenor trombone and bass trombone, and one for alto trombone and tuba. Double reeds included three versions of bassoon and oboe and one of contrabassoon, English horn and Dulcian. Single reeds included three versions of clarinet and one of basset horn, alto saxophone, and tenor saxophone. Finally, bowed strings included two versions of violin, viola, cello and double bass. The class distribution of each test set (*pp* and *ff*) was the following: brass (405, 367), double reeds (296, 304), single reeds (261, 263), and bowed strings (409, 369).

In these experiments (lower panel of Table 2) models trained on MPS scored the highest results on all test sets. The best performances were achieved with all three representations on single reeds instruments, suggesting for this family clear underlying spectrotemporal modulation differences between the two dynamics extremes. When compared to the results on the violin set in the previous experiment, the low accuracies obtained on the violin family can be interpreted as a sign of overfitting. Generally, models trained on MPS appear to be more robust against selection bias.

3.3 10-fold Cross Validation

To offer a third comparison where selection bias is avoided, a 10-fold cross validation was performed on the entire randomized dataset. All input representations delivered high performances (Table 3). These results showed that there is no significant difference between models trained on MPS and ERB-scaled spectrograms, but that when selection bias is taken into account and prevented, mel-spectrograms provide slightly more discriminative features to the CNN.

4. AN MPS-BASED TIMBRE DESCRIPTOR

4.1 Definition

Model Visualization Through Saliency Maps

Saliency maps indicate the input regions whose change is most relevant for maximizing the output of a chosen neuron, in this case the two class-specific neurons of the final layer of the CNN. Saliency maps can be interpreted as a correlation maps, where a salient regions basically represents an area of the input that highly correlates with the chosen target. All maps were obtained with the function *visualize_saliency* of the keras-vis toolkit [19]. The average of the maps relative to all inputs of a specific class was calculated for each family separately.

Comparing the MPS saliency maps of all analysed families revealed a clearer pattern (see Fig. 2). For both classes the most salient region appears to be generally located along the spectral modulation axis and near the origin (i.e., low temporal modulations), while in the fortissimo maps of brass and double reeds the salient region appears to be slightly shifted toward higher temporal modulations. What is surprising is that for all instrument families, in at least one of the two averaged saliency maps, a similar well-defined central salient region is present. This suggests that in the low-temporal modulations of the MPS is somehow precisely encoded the dynamic strength of single notes of sustained instruments.

In order to investigate to what extent this area is alone effective in predicting the dynamics of single notes, a scalar timbre descriptor is here defined as the average computed from 2 to 16 cycles/kHz of the sum of the temporal frames (from 0 to 3 Hz) of the MPS:

$$\frac{1}{K_{16} - K_2} \cdot \sum_{k=K_2}^{K_{16}} \sum_{j=J_0}^{J_3} MPS(j,k)$$

where K_2, K_{16} and J_0, J_3 are the spectral and temporal indices for 2 and 16 cycles/kHz and 0 and 3 Hz respectively. From now on, the descriptor will be referred to as steady spectral modulation power (*steady SMP*).

4.2 Baseline and Evaluation

Baseline

The following audio features were identified by Weinzierl et al. [4] as good discriminators for musical dynamics (within and across instruments) and will be used as a baseline to compare the discriminative power of the steady SMP. As previously mentioned, the attack phase of the note was not considered in the analysis, so the final features set consists of only spectral flatness (STFTmag, median) and skewness (ERBfft, median). Both descriptors were extracted in MATLAB using the default parameters of the TTB.

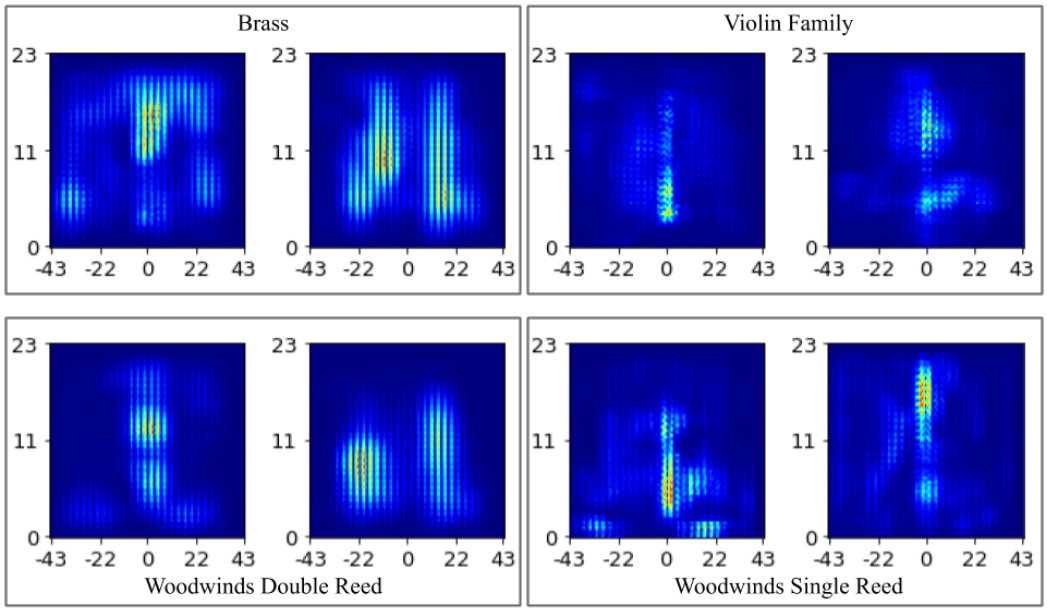

Figure 2: Averaged MPS saliency maps of different instrument families (four pairs, *pp* and *ff*). The ordinate is in Hz, and given that the mel-scale is not directly convertible in octaves, the abscissa is shown in cycles/kHz with only the extremes and a middle point for reference.

Dataset	**steady SMP**	**spec. flat. & skew.**
Brass	0.84	**0.89**
Double reeds	0.80	**0.82**
Flute	**0.74**	0.56
Single reeds	**0.94**	0.84
Violin family	0.84	**0.85**

Table 4: Family-wise LDA F1 scores for the steady SMP against spectral flattness and skewness.

Experiments

In order to determine the behaviour of the descriptor on different timbre families, a 10-fold cross validated LDA was performed on the steady SMP and on the two TTB descriptors, where this time, each timbre family was treated separately as an independent dataset. Table 4 shows that the steady SMP showed good predictive power for the dynamics of single notes in all instrument families. While the two TTB descriptors scored better on three different families (brass, double reeds and violin family), the difference in performance is never too important, but on flute and single reeds, where the steady SMP has totalled increments of circa +20% on the flute set and +10% for single reeds. Noteworthy is the behaviour on the flutes, where if compared to previous results one would have expected a weaker discriminative power.

Figure 3 shows the distribution of SMP values of *pp* and *ff* notes for the entire dataset and the flute set. For both classes in the flute set, SMP values fall under 3×10^8, which corresponds to the threshold between *pp* and *ff* notes for the entire dataset. This could help to explain why in previous experiments all classifiers showed worse results on the flute set. Finally, when tested against spectral skewness and flatness in a 10-fold cross validated LDA on the entire randomized dataset (Table 5), the SMP was able to gain a 44.8% error reduction (Wilcoxon p = 0.002).

Features set	**F1 score**
steady SMP	$\mathbf{0.84 \pm 0.016}$
spec. flat. & skew.	0.71 ± 0.022

Table 5: Results of the 10-fold cross validated LDA for steady SMP and spectral flattness and skewness.

5. CONCLUSIONS

A convolutional neural network for musical dynamics classification was introduced, and the use of modulation power spectra was proposed and tested against two well-established time-frequency representations as input for the system. The aim of the study was to investigate relevant spectrotemporal modulation patterns in single notes of instruments played at the two dynamics extremes, pianissimo and fortissimo.

In the first and third experiments, where the test sets were in different ways representative of the training data, the models trained on the two time-frequency representations scored better than models trained on the MPS. This could be explained in light of specific relations between musical dynamics and pitch height that should have been better captured by time-frequency representations. Notably, apart for aerophone instruments, the models trained on the MPS scored better than the others when they were tested for robustness (second experiment), in a task that promoted

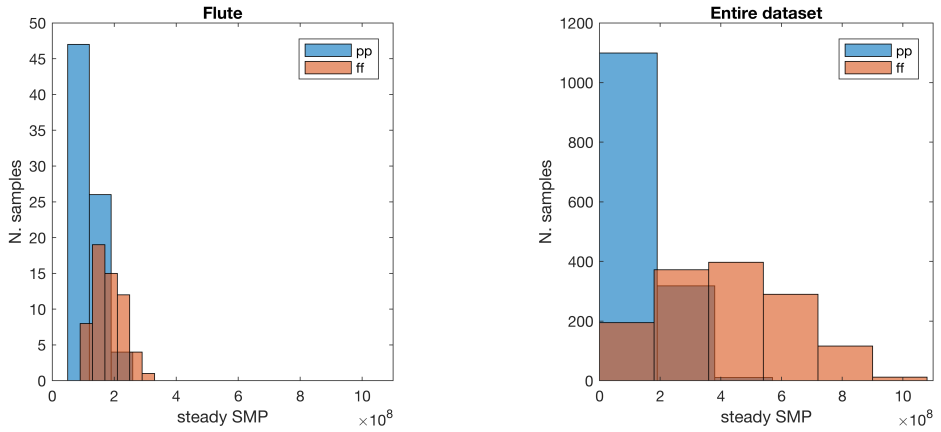

Figure 3: Histograms showing the distribution of the steady SMP in the two classes.

selection bias. These remarkable results can be attributed to the pitch invariance afforded by the MPS, that facilitated the network in the extraction of features that would generalize better on unrepresented data.

Visual inspection suggested that the most salient region for dynamics recognition is located among low temporal modulations. Consequently, as proposed in [10], a MPS-based scalar descriptor was implemented and tested. Compared against two descriptors of timbre selected for the same task (and on the same dataset) [4], the proposed MPS-based descriptor showed comparable, when not better predictive power than the two combined spectral descriptors. Particularly, when compared with the baseline in a 10-fold cross validated LDA, the proposed descriptor obtained 44.8% error reduction.

Clear differences in the modulation space of notes played at different musical dynamics are encoded in low temporal modulations. Yet it should be noted that all the experiments were performed on single notes of sustained instrument played at only two dynamics extremes (*pp* and *ff*), hence limiting the generalizability of these findings. Future studies should expand on the results by analysing impulsive sounds (such as plucked strings and piano notes) and by including different dynamic gradations (i.e., *p*, *mp*, *mf* and *f*). Moreover in future works, the combination of MPS and CNN should be also tested on common MIR tasks, such as instrument and genre recognition.

Deep learning methods made the use of hand-crafted features obsolete at the expense of interpretability of the automatically extracted features, which are mostly too complex and task-specific. This study showed that the translation invariance afforded by modulation power spectra promotes the extraction of features that are more robust and generalizable, while at the same time more locally spaced in the input domain than those based on time-frequency representations. This property of the MPS and its invertibility, combined with visualization tools for neural networks, could facilitate the investigation of such features from a perceptual point of view.

Acknowledgments

We thank Andreas Schuller and Philipp Scholze for assisting with earlier versions of the manuscript.

6. REFERENCES

[1] K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay, Eds., *Timbre: Acoustics, Perception, and Cognition*. Cham: Springer, 2019.

[2] T. Nakamura, "The communication of dynamics between musicians and listeners through musical performance," *Perception & Psychophysics*, vol. 41, no. 6, pp. 525–533, 1987.

[3] M. Fabiani and A. Friberg, "Influence of pitch, loudness, and timbre on the perception of instrument dynamics," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. EL193–EL199, 2011.

[4] 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," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1347–1355, 2018.

[5] M. Caetano, C. Saitis, and K. Siedenburg, "Audio content descriptors of timbre," in *Timbre: Acoustics, Perception, and Cognition*, K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay, Eds. Cham: Springer, 2019.

[6] J. Schlüter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 6979–6983.

[7] Y. Han, J. Kim, K. Lee, Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 1, pp. 208–221, 2017.

[8] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," *arXiv preprint arXiv:1604.06338*, 2016.

[9] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS computational biology*, vol. 5, no. 3, p. e1000302, 2009.

[10] E. Thoret, P. Depalle, and S. McAdams, "Perceptually salient regions of the modulation power spectrum for musical instrument identification," *Frontiers in Psychology*, vol. 8, p. 587, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2017.00587

[11] S. Shamma, "On the role of space and time in auditory processing," *Trends in cognitive sciences*, vol. 5, no. 8, pp. 340–348, 2001.

[12] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: the biological bases of musical timbre perception," *PLoS computational biology*, vol. 8, no. 11, p. e1002759, 2012.

[13] T. M. Elliott, L. S. Hamilton, and F. E. Theunissen, "Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 389–404, 2013.

[14] E. Thoret, P. Depalle, and S. McAdams, "Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments," *The Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. EL478–EL483, 2016.

[15] S. Weinzierl, M. Vorländer, G. Behler, F. Brinkmann, H. von Coler, E. Detzner, J. Krämer, A. Lindau, M. Pollow, F. Schulz *et al.*, "A database of anechoic microphone array measurements of musical instruments," 2017.

[16] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[17] J. Heeris, "Gammatone-based spectrograms, using gammatone filterbanks or fourier transform weightings," https://github.com/detly/gammatone, Oct 2018.

[18] M. Slaney, "Auditory toolbox," *Interval Research Corporation, Tech. Rep*, vol. 10, no. 1998, 1998.

[19] R. Kotikalapudi and contributors, "keras-vis," https://github.com/raghakot/keras-vis, 2017.