AUTOMATIC MUSIC TAGGING WITH HARMONIC CNN

Minz Won Universitat Pompeu Fabra minz.won@upf.edu
Sanghyuk Chun Clova AI Research sanghyuk.c@navercorp.com
Oriol Nieto Pandora Media Inc. onieto@pandora.com
Xavier Serra Universitat Pompeu Fabra xavier.serra@upf.edu

EXTENDED ABSTRACT

In this paper, we introduce the Harmonic Convolutional Neural Network (Harmonic CNN), a music representation model that exploits the inherent harmonic structure of audio signals. The proposed model outperforms previous approaches in automatic music tagging on the MagnaTagATune (MTAT) dataset [5]— see Table 1.

Feature design was one of the main focuses in early stages of music informatics research (MIR), where such features were later used as input to machine learning models to, e.g., bridge the semantic gap [2] between signal-level features and high-level music semantics. However, with the emergence of deep learning, recent MIR models can learn feature representations in an end-to-end data-driven way. Hence, minimum domain knowledge is required in the preprocessing step (i.e., short-time Fourier transform). Recent works, such as sample-level CNN [6], use raw audio waveforms directly as their inputs. With no domain knowledge in its architecture design and preprocessing, sample-level CNN yielded state-of-the-art results in music tagging [4].

Nevertheless, we believe domain knowledge may facilitate efficient representation model design, especially when a limited amount of data is available [7]. To this end, we propose the Harmonic CNN based on empirical evidence that the harmonic structure plays a key role in human auditory perception [9].

The first layer of Harmonic CNN, the Harmonic convolution layer, learns a band-pass filterbank using sinc functions as proposed in the SincNet [8]. Since front-end filters of a standard CNN perform similar to a band-pass filterbank, adopting SincNet architecture is a reasonable choice. The original implementation of SincNet learns two cut-off frequencies, f_k^{low} and f_k^{high}, to form a k-th band-pass filter[1]. However, based on the empirical research, we can approximate the bandwidth with a linear transformation of the center frequency [3] so that the band-pass filter can be expressed by a function of center frequency [2]. The bandwidth becomes wider as the center frequency increases — see Figure 1c. Thus, only the center frequencies f_k and bandwidth parameters α, β, and Q are learnable in the Harmonic convolution layer. We initialize the center frequencies with quarter tone intervals: $f_k = f_{min} \cdot 2^{k/24}$; where the lowest frequency f_{min} is C1 (32.7 Hz). Bandwidth parameters α and β are initialized with the values of equivalent rectangular bandwidth (ERB) which are 0.1079 and 24.7, respectively [3].

We force the Harmonic convolution layer to capture harmonic characteristics by stacking harmonic band-pass filters. As shown in Figure 1b, the Harmonic convolution layer automatically generates harmonic band-pass filters from each f_k. The m-th harmonic band-pass filter can be depicted with center frequency $m \cdot f_k$ and bandwidth $(\alpha \cdot m \cdot f_k + \beta)/Q$. As a result, the output tensor of the Harmonic convolution layer has a shape of $B \times H \times F \times T$, where B is batch, H is harmonic, F is frequency, and T is time (Figure 1d). This can be interpreted as a pseudo harmonic CQT [1] with learnable parameters. By regarding harmonics as channels, the Harmonic convolution layer is followed by 2-D CNN (Figure 1a). More elaborated architecture design can be done with this back-end but we simply used a stack of conventional 3×3 convolution filters since our main interest is to show the importance of the harmonic relationship.

For future work, we would like to determine whether the Harmonic CNN outperforms in more diverse tasks beyond music such as speech recognition or acoustic event detection; and if the Harmonic CNN scales accordingly when using large amounts of training data.

[1] $g[t, f_k^{high}, f_k^{low}] = 2f_k^{high} sinc(2\pi f_k^{high} t) - 2f_k^{low} sinc(2\pi f_k^{low} t)$, where $sinc(x) = \sin(x)/x$
[2] $f_k^{high} = f_k + BW_k/2$, $f_k^{low} = f_k - BW_k/2$, where $BW_k = (\alpha f_k + \beta)/Q$

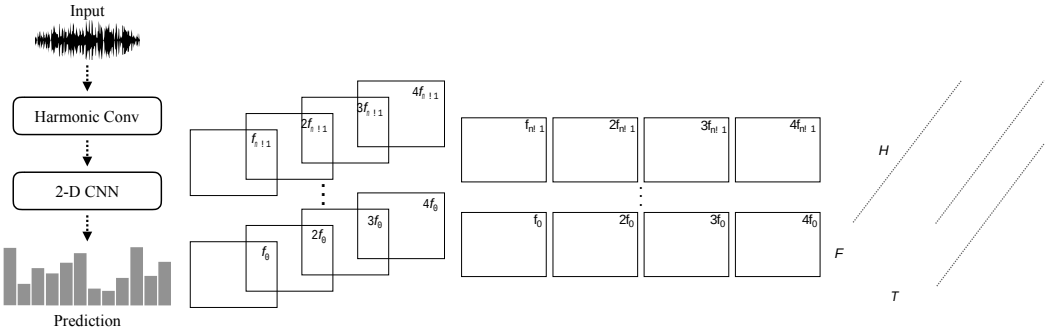

Figure 1: (a) model architecture; (b) harmonic convolution filters in time domain; (c) magnitude frequency response of harmonic convolution filters; (d) an output tensor of the Harmonic convolution layer.

MODELS	AUROC	AUPR
Musicnn [7]*	0.9089	0.4503
Sample-level CNN [6]	0.9054	0.4422
Sample-level CNN + SE [4]	0.9083	0.4500
Sample-level CNN + Res + SE [4]	0.9075	0.4473
Harmonic CNN (proposed)	**0.9146**	**0.4628**
Harmonic CNN + Res (proposed)	**0.9155**	**0.4657**

Table 1: Performance comparison of music tagging models. Reported values are averaged across three runs and (*) indicates a reproduced result since the original paper used a different data split.

ACKNOWLEDGMENTS

This work was funded by the predoctoral grant MDM-2015-0502-17-2 from the Spanish Ministry of Economy and Competitiveness linked to the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

REFERENCES

[1] Rachel M Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan Pablo Bello. Deep salience representations for f0 estimation in polyphonic music. In *ISMIR*, pages 63–70, 2017.

[2] Oscar Celma, Perfecto Herrera, and Xavier Serra. Bridging the music semantic gap. 2006.

[3] Brian R Glasberg and Brian CJ Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2):103–138, 1990.

[4] Taejun Kim, Jongpil Lee, and Juhan Nam. Sample-level cnn architectures for music auto-tagging using raw waveforms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*.

[5] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pages 387–392, 2009.

[6] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. *arXiv preprint arXiv:1703.01789*, 2017.

[7] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. *arXiv preprint arXiv:1711.02520*, 2017.

[8] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE, 2018.

[9] William A Sethares. *Tuning, timbre, spectrum, scale*. Springer Science & Business Media, 2005.