A preliminary evaluation of robustness to noise using the compositional hierarchical model for music information retrieval

Matevž Pesek[1], Aleš Leonardis[1,2], Matija Marolt[1]

[1] *University of Ljubljana, Faculty of computer and information Science*
[2] *Centre for Computational Neuroscience and Cognitive Robotics, School of Computer Science, University of Birmingham*
e-mail: {matevz.pesek, ales.leonardis, matija.marolt}@fri.uni-lj.si

Abstract

This paper presents a preliminary evaluation of a compositional hierarchical model for music information retrieval in terms of its robustness to noise. The model has been previously introduced and evaluated for automated chord estimation and multiple fundamental estimation tasks. The model claims to provide mechanisms aiding the model to perform well on noisy data. We evaluated the model's robustness to noise by performing a set of tests by adding the white and pink noise to the audio data and re-evaluating the model for the automated chord estimation (ACE) task. The paper describes the experiment and shows the model's ability of producing relevant features in suboptimal conditions, resulting in a graceful degradation of the classification accuracy for the ACE task.

1 Introduction

The field of music information retrieval (MIR) only formally exists for a good decade; however, it has reached a significant expansion in tasks and solutions focusing on a variety of music-oriented aspects [2]. These aspects include extraction of high-level music descriptors from music, such as melody, harmony and rhythm, as well as highly perceptual or subjective tasks involving mood estimation and genre recognition. None of these tasks has reached a state of perfect solution. Nonetheless, new approaches are constantly raising the level of the state-of-the-art results.

The deep learning architectures have been known for some time, yet only recently have such approaches been introduced into the MIR field, with its most prominent representative model, the deep belief networks (DBNs). The DBNs represent a single computational model which can and has been applied to a variety of tasks. The concept of deep learning has grown in popularity in the fields of signal processing [14], audio processing [9] and MIR. Lee [6] presented one of the first attempts of using deep belief networks (DBNs) on audio signals, where convolutional DBNs were applied to the speaker identification task. Focusing on music signals, Hamel and Eck [3], evaluated DBNs for genre recognition. The DBNs show great potential for many tasks that involve high-level feature extraction, such as emotion recognition, since there is usually no trivial spectral or temporal feature that could be used to model the high-level representation in question. Schmidt and Kim [13] showed promising results by using a DBN for extraction of emotion-based acoustic features. Overall, recent research has shown great interest and success in using features learned from music signals, in contrast to previously used hand-crafted features. There is a vast expansion of deep learning in MIR to be expected, as anticipated by Humphrey [4].

This paper focuses on another deep learning architecture — a biologically-inspired compositional hierarchical model for music information retrieval [11, 12]. The proposed model poses an alternative to recent deep learning architecture approaches. Its main difference from the latter is in its transparent structure, thus allowing representation and interpretation of the signal's information extracted on different levels. The model has been applied to the MIR tasks of automated chord estimation and multiple fundamental frequency estimation. Although newly introduced, this alternative shows great potential for MIR. We re-evaluate the model for the task of automated chord estimation by applying several scenarios of adding noise to the audio recordings and comparing the model's performance. The paper is structured as follows: the compositional hierarchical model is presented in Section 2, the experiment of evaluating the robustness to noise is explained in Section 3. We elaborate on the results and conclude the paper in Section 4.

2 Compositional Hierarchical Model

The structure of the model is inspired by work in computer vision, specifically the hierarchical compositional model presented by Leonardis and Fidler [7].

Their model represents objects in images in a hierarchical manner, structured in layers from simple to complex image parts. The model is learned from the statistics of natural images and can be employed as a robust statistical engine for object categorization and other computer vision tasks. Since no trivial transformation between visual and audio signal domains exists, it took significant effort to transform the model to the MIR domain.

2.1 Structure

The compositional hierarchical model provides a hierarchical representation of the audio signal, from the signal components on the lowest level, up to individual musical events on the highest levels. The model is built on the belief of the signal's ability of hierarchical decomposition into atomic blocks, denoted as *parts*. According to their complexity, these parts can be structured across several layers from less to the more complex. Parts on higher layers are expressed as compositions of parts on lower layers — similarly as a chord is composed of several pitches, or a pitch represents a composition of several harmonics. A part can therefore describe individual frequencies in a signal, their combinations, as well as pitches, chords and temporal patterns, such as chord progressions.

The first $\mathcal{L}_0$ layer of the model corresponds directly to the time-frequency representation of the music signal, as parts correspond to individual frequency bins and part activations to their magnitudes. Subsequent higher layers $\{\mathcal{L}_1..\mathcal{L}_n\}$ consist of compositions of parts on lower layers and thus represent relative combinations of frequency components in the signal (Fig 1).

The model is built layer-by-layer in an unsupervised manner from a given data set. While building a new layer $\mathcal{L}_n$, a set of possible part compositions is constructed by observing the statistics of co-occurrences of part activations on the $\mathcal{L}_{n-1}$ layer. Frequently co-occurring parts are chosen and combined into new parts. Compositions are denoted as links between parts in Figure 1. Composition i on layer $\mathcal{L}_n$ can be formally defined as a structure containing parts from a layer below: a central part C, and a secondary part S. We name the parts forming a composition subparts. A composition can be defined as:

$$P_{n,i} = \{C_{n-1,j}, S_{n-1,k}, (\mu_{n,i}, \sigma_{n,i})\}, \qquad (1)$$

where $C_{n-1,j}$ and $S_{n-1,k}$ are the central and secondary subparts from layer $n-1$, while $\mu_{n,i}$ and $\sigma_{n,i}$ define a Gaussian limiting the difference between locations of subpart activations. A part $P_{n,i}$ activation is composed of two values: activation location $L_{Pn,i}$, which represents the location (frequency) at which the part is activated, and activation magnitude $A_{Pn,i}$, which represents the strength of activation. The location of part's activation is defined simply as the location of activation of its central subpart:

$$L_{Pn,i} = L_{Cn-1,j}. \qquad (2)$$

Thus, central parts of compositions on different layers propagate their locations upwards through the hierarchy. The magnitude of activation is defined as:

$$A_{Pn,i} = tanh[G(L_{Cn-1,j} - L_{Sn-1,k}, \mu, \sigma) \cdot (A_{Cn-1,j} + A_{Sn-1,k})], \qquad (3)$$

where *tanh* stands for the hyperbolic tangent function that limits the magnitude to [0,1) and G represent the Gaussian that limits the difference in locations of the central part and the subpart according to μ and σ.

However, in order to avoid excessive redundancies and obtain the more significant compositions, a greedy part selection algorithm is used, which maximises the amount of previous-layer activations covered by a newly composed part, while minimizing the number of parts in a layer. In this paper, it is our intention to reproduce the three-layer structure of the model used for the automated chord estimation task, described in [12].

When a model is built, it can be used for inference over any desired data set. Part activations across all layers of the hierarchy are calculated for each time-frame of the time-frequency representation of the analysed music signal. These activations can be used as audio descriptors for tackling various MIR tasks. A more thorough explanation of the model's structure is also provided in [10].

2.2 Relativity and shareability of parts

We point out the relativity and shareability of parts as two key differences of the model, when compared to the DBN approach. As shown in Figure 1, one part can be activated at multiple locations. On $\mathcal{L}_1$ layer, parts produce activations for all co-occurrences of subparts, not defined by their position on an absolute spectral scale, but rather by the offset of locations between co-occurring subparts. Thus, $P_{2,2}$ part, with subparts positioned in an offset of one octave, is activated on a set of locations $\{294Hz, 440Hz\}$ at given time-frame t_1. The $\mathcal{L}_0$ layer and $\mathcal{L}_1$ layer can be thus observed as a fully connected sub-graph where each pair, consisting of a $\mathcal{L}_1$ and a $\mathcal{L}_0$ part are connected. The necessary condition for each activation in the set of $P_{1,i}$ part

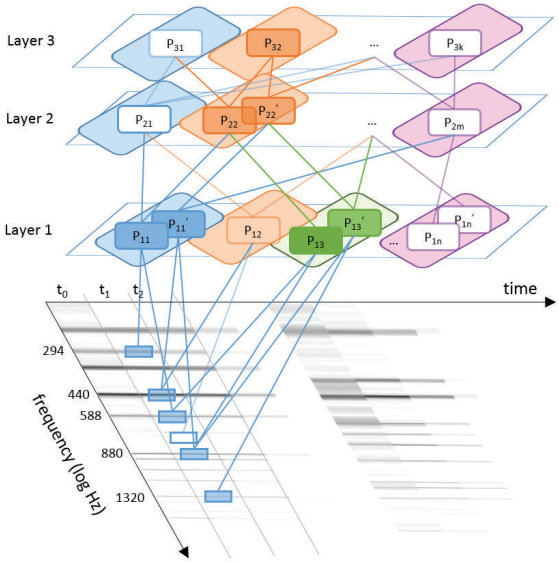

Figure 1: An abstraction of the compositional hierarchical model. Parts on the input layer correspond to signal components in the time-frequency representation. Parts on higher layers are compositions of lower-layer parts (denoted as links in the figure). A part may be contained in several compositions, e.g. P_{11} on the first layer is part of compositions P_{21}, P_{22} and P_{2m} on the second layer. Several depictions of the same part (e.g. part instances P_{11} and P'_{11}) denote several activations of the part on different locations (all instances of a part on a layer are marked with the same outlined color). Parts activated in t_1 are shown filled with color.

activations $A_{P1,i}$ at given time-frame are two activations of $\mathcal{L}_0$ parts, co-occurring with the given offset. Thus, $\mathcal{L}_1$ parts are *relative* by the definition their activations.

The relativity of parts is retained for higher-layer parts $\{\mathcal{L}_0..\mathcal{L}_N\}$. However, the $\mathcal{L}_n$ parts connect only to their composing $\mathcal{L}_{n-1}$ subparts. The part $P_{n,i}$ activations retain the property of the location which is propagated through the $C_{n-1,j}$ subpart's location of the activation. Both $C_{n-1,j}$ and $S_{n-1,k}$ subparts may possess a set of activations. The offset of $P_{n,i}$ activation provides the eligibility criterion for $P_{n,i}$ part activation. The relativity of parts enables the model to provide a single part representing an abstract high-level concept regardless of its location of presence in the observed signal. For example, a single part can represent a concept of a pitch, moreover, the same part, activated on two different locations, may compose a part representing an interval on the consequent layer.

The feature of the parts' relativity provides the possibility of one part covering all present representations reflecting similar structure, regardless the location. The shareability of parts is reflected when observing the connections between consecutive $\mathcal{L}_{n-1}$ and $\mathcal{L}_n$ layers of the model. Any single $\mathcal{L}_{n-1}$ part may form several $\mathcal{L}_n$ compositions, thus eliminating the need of retaining several instances of a single abstract representation for each composition.

3 Evaluation of the model

The experiment was performed using a three-layer compositional model by exporting the (third) octave-invariant features, similar to the chroma features. We evaluated the model on the first two albums of *The Beatles* dataset, provided by C. Harte. We added different amounts of pink and white noise to the original audio recordings with an intention to observe the degradation of the classification accuracy of the automated chord estimation task.

There are several evaluations of approaches when considering the robustness to noise. Boulanger-Lewandowski et al. [1] provided an evaluation to a set of noise-types with variable signal-to-noise ratio. Lardeur et al. [5] evaluated the possible use of prior knowledge for robustness by adding the equalization, reverberation and compression effects. No strict noise-robustness evaluation procedure has yet become a standard in the MIR field. Nonetheless, Mauch and Ewert developed [8] an Audio Degradation toolbox (ADT) designed especially for such evaluation. The ADT offers a variety of audio effects such as noise, reverberation and other effects resembling real situations. For this preliminary evaluation we used a subset of the provided toolbox effects for the experiment and generated noisy audio input using pink and white noise with SNR between [20, 0] dB with a step of 5 dB, similar to [1] and others.

We performed the experiment as follows. The hierarchical model was trained on 88 piano key strokes. For the experiment, we used the albums *Please please me* (denoted as *Album1*) and *With the Beatles* (denoted as *Album2*). The audio files of first two albums were reproduced using both pink and white noise at given signal-to-noise ratios (total of 280 noisy audio files and 28 originals). Using the trained hierarchical model, we produced octave-invariant features for all audio files, including the originals.

We performed the procedure of classifying the output using a Hidden Markov model (HMM). The HMM was trained on the clean features, calculated on the original (noise-free) albums and tested on the noisy albums exclusively; for example, we trained the HMM on noise-free *Album1* and classified all the noisy versions of *Album2*. We repeated the process by switching the *Album2* as the train set and noisy versions of *Album1* for the test set. Figure 2 shows the graceful degradation of the classification accuracy, reflecting the impact of the added pink and white noise to the data. Moreover, the model seems

to be less severely affected by the pink noise, compared to the degradation using the white noise.

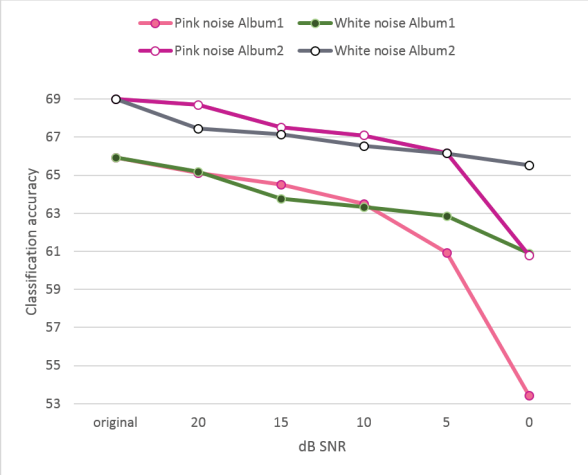

Figure 2: A graph representing the degradation of the automated chord estimation task classification accuracy for the first two albums of the *the Beatles* compilation. Values marked with *Album1* represent the results of classifying the first album using a HMM learned on the second album, *Album2* represents the results by learning on the first and testing on the second album. The graph shows a graceful degradation. The model is more affected by the white noise than to the pink noise.

4 Conclusion

This paper presented a preliminary evaluation of the compositional hierarchical model's robustness to noise. The models appears to be a valuable alternative to other deep architecture approaches. We further confirmed this hypothesis by evaluating the model on audio chord estimation task, extending it to include a variety of different noisy audio data and observing the output. By the graceful degradation of the classification accuracy over a set of noise parameters including two noise types added at various SNRs, we can conclude the model appears to perform excellently in the given conditions. It is yet to be confirmed this robustness may be a result of the model's training set, as shown by [5]. We plan on performing an extensive evaluation adding a variety of affects, including the ones described in [5], and expanding the evaluation to other tasks.

References

[1] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. High-dimensional sequence transduction. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3178–3182. IEEE, May 2013.

[2] J. Stephen Downie, Andreas F. Ehmann, Mert Bay, and M. Cameron Jones. The Music Information Retrieval Evaluation eXchange: Some Observations and Insights. In Wieczorkowska A.A. and Ras Z.W., editors, *Advances in Music Information Retrieval*, pages 93–115. Springer-Verlag, Berlin, 2010.

[3] Philippe Hamel and Douglas Eck. Learning Features from Music Audio with Deep Belief Networks. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 339–344, 2010.

[4] Eric J. Humphrey, Juan P. Bello, and Yann LeCun. Moving beyond feature design: deep architectures and automatic feature learning in music informatics. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Porto, 2012.

[5] M Lardeur, S Essid, G Richard, M Haller, and T Sikora. Incorporating prior knowledge on the digital media creation process into audio classifiers. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1653–1656, April 2009.

[6] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems*, pages 1096–1104, 2009.

[7] Aleš Leonardis and Sanja Fidler. Towards scalable representations of object categories: Learning a hierarchy of parts. *Computer Vision and Pattern Recognition, IEEE*, pages 1–8, 2007.

[8] Matthias Mauch and Sebastian Ewert. The Audio Degradation Toolbox and its Application to Robustness Evaluation. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, pages 83–88, 2013.

[9] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton. Acoustic Modeling using Deep Belief Networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):14–22, 2010.

[10] Matevž Pesek, Aleš Leonardis, and Matija Marolt. A compositional hierarchical model for music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Taipei, 2014.

[11] Matevž Pesek and Matija Marolt. Chord estimation using compositional hierarchical model. In *6th International Workshop on Machine Learning and Music, held in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD 2013*, 2013.

[12] Matevz Pesek, Ales Leonardis, and Matija Marolt. Boosting audio chord estimation using multiple classifiers. In *International Conference on Systems, Signals and Image Processing (IWSSIP), 2014*, pages 107–110, Dubrovnik, 2014. IEEE.

[13] Erik M. Schmidt and Youngmoo E. Kim. Learning emotion-based acoustic features with deep belief networks. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 65–68. IEEE, October 2011.

[14] Dong Yu and Li Deng. Deep Learning and Its Applications to Signal and Information Processing [Exploratory DSP. *IEEE Signal Processing Magazine*, 28(1):145–154, January 2011.