Boosting Audio Chord Estimation using Multiple Classifiers

Matevž Pesek 1, Aleš Leonardis 2, Matija Marolt 1

1 Laboratory for computer graphics and multimedia
Faculty of Computer and Information Science University of Ljubljana, Slovenia
2 Centre for Computational Neuroscience and Cognitive Robotics School of Computer Science
University of Birmingham, United Kingdom
matevz.pesek@fri.uni-lj.si, ales.leonardis@fri.uni-lj.si, matija.marolt@fri.uni-lj.si

***Abstract*—The paper addresses the task of automatic audio chord estimation using stacked generalization of multiple classifiers over Hidden Markov model (HMM) estimators. We evaluated two feature types for chord estimation: a new compositional hierarchical model and standard chroma feature vectors. The compositional hierarchical model is presented as an alternative deep learning approach.**

Both feature types are further modelled with two separate Hidden Markov models (HMMs) in order to estimate chords in music recordings. Further, a binary decision tree and support vector machine are proposed binding the HMM estimations into a new feature vector. The additional stacking of the classifiers provides a classification boost by 17.55% with a binary decision tree and and 21.96% using the support vector machine.

***Keywords*—compositional hierarchical model, deep learning, stacking generalization, audio chord estimation**

I. Introduction

The field of music information retrieval (MIR) undertook significant expansion in tasks and solutions in the short timespan of its existence ([1], [2]). The tasks include extraction of high-level music descriptors such as melody estimation, chord estimation and beat tracking, as well as highly perceptual tasks involving mood estimation, genre recognition and artist influence. Solutions have not come to a perfect one for any of the described task yet, however numerous approaches proposed each year are improving the state-of-the-art rapidly. Recently, deep belief networks as an alternate single model for a variety of tasks have been successfully introduced to the field.

Audio chord estimation is a well known music information retrieval task with the goal of transcribing the succession of chords in an audio signal [3]. The task is not uniquely defined, as several interpretations of chords, as well as their boundaries may exist.

In the paper, we present a new deep architecture which may be used for chord estimation and other MIR tasks. We propose a compositional hierarchical model as a white-box representation of the musical building blocks contained in analyzed audio signal. We evaluate the usefulness of the model for audio chord estimation.

The remainder of this paper is structured as follows: Section 2 explains the novelty of the proposed model and the methods developed. Section 3 includes evaluation of the model and section 4 concludes the paper and provides guidelines for future development.

II. The compositional hierarchical model

Our model is built on the assumption that a complex signal can be decomposed into several constituent building blocks - parts. These parts exist at various levels of granularity and represent sets of entities describing the signal. Considering previous research, we find such hierarchical modularity logical [4]. According to their complexity, parts can be structured across several layers from less to the more complex. Parts on higher layers can be expressed as compositions of parts on lower layers (e.g.: a chord is composed of several pitches). A part can thus describe individual frequencies in a signal, their co-occurrences, as well as pitches, chords and complex temporal patterns, such as chord progressions.

Our compositional hierarchical model (CHM) is a realisation of the described structure. It is a multi-layered model where each layer consists of several parts. The model is built bottom-up layer-by-layer and does not contain hidden layers - it is thus a white-box structure. A similar model has previously been presented in the field of computer vision ([5], [6]) giving excellent results.

A part in our model represents a fragment of information representing the underlying signal. Each layer bases on the set of parts from previous layer combining them into ”compositions”, forming the entities of the new layer. Thus, a part is a composition of two or more subparts on a lower layer. The activation of a part on i-th layer $P_{i,a}$ at given time i directly depends on the activations of its subparts forming the subset $S_{P_\bullet}$:

$$Act(t, P_{i,a}) = \forall P_\alpha \in S_{P_\bullet} : \sum Act(t, P_\alpha) \qquad (1)$$

The power of the activation is proportional to the amount of information covered in the signal. An abstract structure of the model is shown in Figure 1.

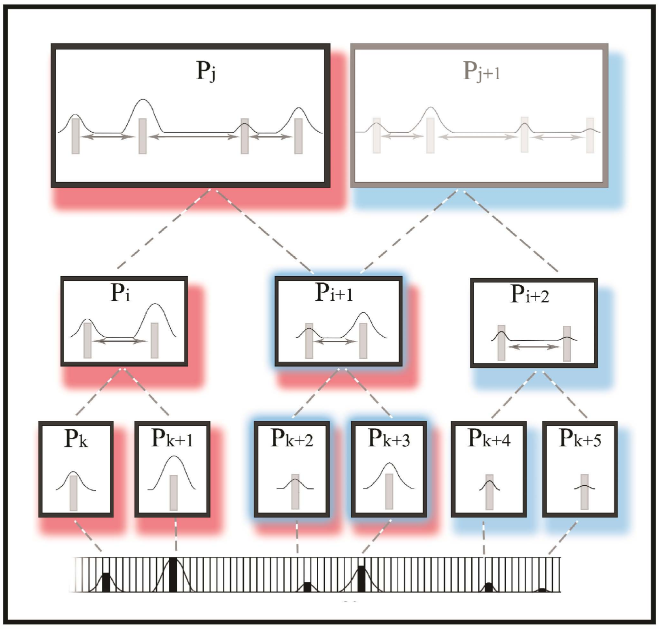

Fig. 1. The inference of the model over A_0, $L_{r/0}$ and $L_{r/1}$ layer. Activation of a part may be formed on several locations at once. Lines combining the activations represent two different locations of activations resulting in two parts using same composition (P_j and P_{j+1}).

The following subsections describe the building phase of the compositional model and elaborate on the biologically-inspired methods.

A. Preprocessing

The preprocessing begins with the sound filtering, emulating the outer and middle ear processing [7]. The process continues with the Constant Q transform [8] with 48 bins per octave, between 55 and 8000 Hz. The step size is 50 milliseconds with the maximum size of Hamming window of 100 milliseconds. We perform additional noise removal by ignoring bins with magnitude lower than -35dB from the peak and -70dB globally [9]. At the end of inference the psychoacoustic masking is performed.

The result of this step is a matrix of $n \times d$ where *n* represents number of frames from the audio and *d = 345* represents number of bins produced by the Constant Q transform. We denote the preprocessing result as the A_0 layer.

B. Building the layers

We define a new layer L_i by introducing a set of possible candidates $\mathbf{C_i}$. We propose a bottom-up approach for feature extraction. The set of possible candidates is populated with possible compositions of parts from lower L_{i-1} layer. The compositions represent the L_{i-1} parts, which co-occurred (simultaneously activated) while inferring the audio signal. A prerequisite for composition of given parts P_a and P_b is a co-occurrence of parts' activations at given time *t*:

$$[P_a, P_b] \in \mathbf{C_i} \equiv \exists t : \{Act(t, P_a) \wedge Act(t, P_b)\} \qquad (2)$$

The set of candidates is further refined to a subset S_i. Due to loose condition for acceptance, there are several parts covering redundant information and random noise. We propose a greedy approach for refining the set. The candidate-picking algorithm performs a trade-off subset extraction, covering the most information with the smallest possible set of parts. This is done by comparing the added amount of information of each candidate over time. Thus, a trade-off criteria are proposed for subset S_i minimization and signal coverage (transformed into activation coverage over time *t*).

$$S_i \subset C_i : min(|S_i|) \wedge max(\frac{\sum_t Act(t, P_a) : \forall P_a \in S_i}{\sum_t Act(t, P_\alpha) : \forall P_\alpha \in C_i}) \quad (3)$$

The trade-off criteria define the robustness of the model. The tendency to keep the number of parts on each layer smaller is a counter part to the possible over-fitting by taking all candidates to the next layer.

C. Inducing non-linear mechanisms

The structure of complex, high-layer parts is often reflected by a portion of information in the signal, due to limited or missing information. In order to extend the robustness of the model, we propose two biologically-inspired mechanisms introducing non-linearity of activations. The missing information in the signal can be partially replaced, depending on the model's knowledge, extrapolated from the learning set. In this case, we allow activations of parts most fittingly covering the information present. The structure fragment which are not reflecting the actual state in the signal are ”hallucinated”. Thus, the higher-level activation appears as if the whole information is present on lower layers. This allows the model to produce hypotheses in situations with no straight result. The hallucination allows us to perform content-dependent perception. A fragment in a time-frame can be covered by a single part. Depending on the information distribution of the time-frame, a similar fragment can be covered as a combination of several other parts. Hallucination process boosts these alternative representations. Therefore, the hallucination process produces multiple hypotheses resulting in a number of robust representations on higher layers of the model.

Although the candidate-picking method penalises the parts redundantly covering the signal, there is additional redundancy induced by the hallucination process. In order to perform hypothesis refinement, an inhibition process is used. The inhibition reduces the number of weaker activations. This limits the number of hypotheses produced by hallucination. Another effect of inhibition is reducing the activations produced by the chance of noise present in the signal. These activations are often of low magnitude. The inhibition process removes the lowest activations depending on the magnitude of the strongest hypotheses provided in the time-frame.

D. Inference

Once the model is built, it can be used on a desired dataset. The model provides a detailed representation of audio signal forming activations on each layer. These activations can be exported for further information extraction using machine-learning algorithms. The activations are calculated for each time-frame in the audio signal, layer-by-layer in a bottom-up manner. A L_i part can be activated on any location in spectrum. The magnitude of activations is proportional to the magnitudes activations of lower L_{i-1} layer part activations composing the L_i part. Thus, there is a possibility of multiple activations of a single part on several locations within a time-frame. Hallucination and inhibition processes are performed on each layer boosting probable hypotheses while inhibiting weak part activations.

The output of the model is a sorted representation of activations per frame. Although all the layers of the model can be outputted, we perform further evaluation on the last two layers of the model.

III. Evaluation

We performed the calculation of chroma feature vectors on the data set consisting of twelve *The Beatles* albums kindly provided by C. Harte. The CHM was built on a set of 88 piano-key recordings, whereas the intermediate-level features were calculated for the Beatles data set.

For each type of features, an ergodic 24-state HMM [10] is built with each hidden state representing a chord from a set of minor and major chords $\{C, ..., B, Cm, ..., Bm\}$. The data set annotations consist of a variety of chords, thus the more complex chords are translated to the root major or minor chord, e.g. C:maj7/E is translated to C:maj. Without information about the starting chords for each song, the a-priori values for matrix π are set to $\frac{1}{24}$ for each hidden state. The initial values for state-transition matrix are based on a doubly-nested circle of fifths, thus providing some musicological know-how based on western music. The parameters for the HMMs are based on previous work done by [11], [12]. The CHM building stage for this experiment is previously described in [13].

A cross-validation for this experiment was performed using one album for the training set while the rest of the data set is being estimated. The Viterbi algorithm is used for the path estimation, thus producing the most probable path through the HMM. Estimated chord sequences are combined for additional classification as shown in Figure 2.

The outputs of both HMMs are combined into a new estimation vector with both estimations as parameters and ground truth. The stacking was performed with a five times two-fold cross validation.

The initial per-frame classification accuracy results are displayed in Table I. The classification accuracy results of both feature types are comparable. In order to evaluate the value of time-dependent evaluation, we have included additional results using support vector machine (SVM). As expected, the results of per-frame SVM are significantly lower, compared to those produced by both HMMs.

The comparability of the results does not automatically imply the ability to boost results by combining them due to the minimal difference between the results, which can reflect the similarity of both feature types. Nevertheless, the stacking process including a binary decision tree (BDT)and SVM was performed. Our hypothesis is formed on an assumption that two different approaches include different interpretations, each with an ability to distinguish some chords better than others. However, combining the interpretations will provide best of each estimation. The stacked BDT and SVM classifiers were chosen in order to perform on the ACE domain which consists of non-linear relations between classes.

With stacking we have improved the classification by 17.55 percent on average to 67.28% using BDT classifier. The improvement using SVM classifier boosted the classification accuracy even further to 71.69 percent. Despite the comparable results of both HMMs and the similarity of both feature types the accuracy boost can be explained by the unsupervised CHM feature learning. The L_{r_1} parts model the pitch features learned in an unsupervised manner. Therefore, the chroma vector representation directly representing pitch classes can

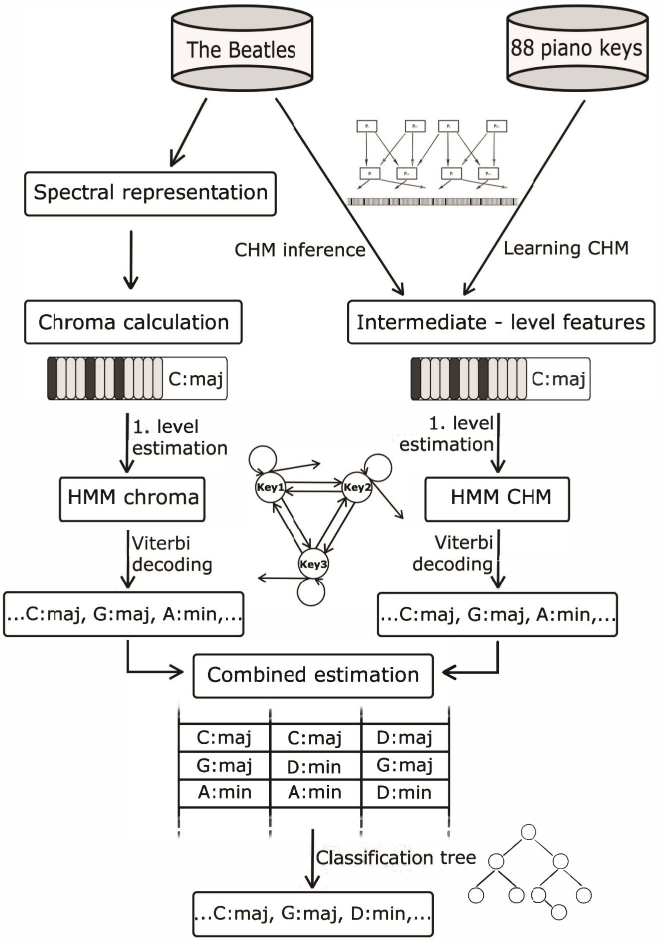

Fig. 2. The outline of the experiment presented in this paper. The CHM is built on 88 piano-keys database and inferred by *The Beatles* dataset. The intermediate-level features are exported and provided as an input to a HMM. The chord progressions are estimated with a Viterbi algorithm. The chord estimations of both HMMs are later joined into a new vector. The estimation is refined by binary decision tree and support vector machine.

TABLE I

The average classification accuracy of HMM_{chm} and HMM_{chroma} models with standard deviances. The classification accuracy of both approaches differs minimally. The per-frame SVM classification accuracy is significantly lower due to the time-independent classification of frame vectors using HMMs. The classification accuracy of both, binary decision tree and support vector machine is significantly greater.

Model	$\overline{CA}$	σ
HMM_{chm}	50.43 %	0.1371
HMM_{chroma}	49.03 %	0.2140
$SVM_{perFrame}$	37.61 %	0.0554
$SVM_{stacking}$	71.69%	
$BDT_{stacking}$	67.28%	

significantly differ from the CHM feature vector.

IV. Conclusion

We proposed the stacking of two distinct approaches for audio chord estimation in order to achieve better classification results. We performed the stacking using the binary decision tree classifier. The compositional hierarchical model was presented as a novel deep architecture approach. The CHM offers clearer insight to the deep architecture modelling demonstrating great potential in the MIR field. The model shares the ability of generating features provided by unsupervised learning with the DBNs. The model also integrates the biologically-inspired methods of hallucination and inhibition, which reflect human auditory perception to some degree.

The CHM is general and can be used for multiple MIR tasks. We evaluated the model for chord estimation, where we compared it to chroma features under a hypothesis of the CHM feature sufficiency evaluated by comparison to chroma features. A small set of 88 piano keys provided as a learning basis for CHM was proven sufficient in order to learn the form of a pitch and sufficient for generating the intermediate-level feature set for chord estimation. Based on the similar classification accuracy results of both approaches the CHM incorporates similar amount of information in the intermediate-level features, whereby unsupervised learning generates these features and no musicological knowledge was hard-coded into the CHM.

We have demonstrated the effectiveness of combining the chord estimations based on different feature types as one of the possibility of improving the classification accuracy for the ACE task. The initial HMM results were significantly boosted by stacking the chroma and CHM features. We extracted these features HMM estimations and with an additional classification of joint estimations with a BDT and SVM.

We intend to research the human temporal perception and its efficient generative implementation. Therefore, we intend to eliminate the need for HMM chord-transition modelling. Building a hierarchical model upon temporal compositions will provide a hierarchy of abstract progression patterns. The temporal hierarchy can be adjusted for beat tracking and tempo extraction tasks. Segmentation, based on the temporal hierarchy, can extend per-frame chord estimation; consequently, achieving far better classification accuracy. Another research interest is the learning of high-level features and developing the model's ability of producing complex compositions representing intervals and harmonies. The ability to perform in other MIR tasks will be implemented in a form of multi-pitch detection with an adjustment of the proposed harmony detection approach.

References

[1] J. Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, no. 1, pp. 2–10, 1999.

[2] N. Orio, "Music Retrieval: A Tutorial and Review," *Foundations and Trends® in Information Retrieval*, vol. 1, no. 1, pp. 1–90, 2006.

[3] G. Peeters and J. Pauwels, "Evaluating Automatically Estimated Chord Sequences," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 749 – 753.

[4] I. Peretz and M. Coltheart, "Modularity of music processing," *Nature Neuroscience*, vol. 6, no. 7, pp. 688–691, 2003.

[5] A. Leonardis and S. Fidler, "Towards scalable representations of object categories: Learning a hierarchy of parts," *Computer Vision and Pattern Recognition, IEEE*, pp. 1–8, 2007.

[6] S. Fidler, M. Boben, and A. Leonardis, "Learning Hierarchical Compositional Representations of Object Structure," in *Object Categorization: Computer and Human Vision Perspectives*. Cambridge University Press, 2009, pp. 196–215.

[7] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models*, 2007.

[8] J. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[9] S. A. Gelfand, *Hearing: An introduction to psychological and physiological acoustics*, 2004.

[10] L. R. Rabiner and B. Gold, *Theory and application of digital signal processing*. Prentice Hall, 1975.

[11] J. P. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals," in *Proceedings of ISMIR*, London, 2005.

[12] H. Papadopoulos and G. Peeters, "Large-case Study of Chord Estimation Algorithms Based on Chroma Representation and HMM," *Content-Based Multimedia Indexing*, vol. 53-60, 2007.

[13] M. Pesek and F. Mihelič, "Hidden Markov model for chord estimation using compositional hierarchical model features," in *Zbornik dvaindvajsete mednarodne Elektrotehniške in računalniške konference*, 2013, pp. 145–148.