Towards CNN-based Acoustic Modeling of Seventh Chords for Automatic Chord Recognition

Christon-Ragavan Nadar[1], **Jakob Abeßer**[1], **Sascha Grollmisch**[1,2]
[1]Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany
[2]Institute of Media Technology, Technische Universität Ilmenau, Ilmenau, Germany
jakob.abesser@idmt.fraunhofer.de

ABSTRACT

In this paper, we build upon a recently proposed deep convolutional neural network architecture for automatic chord recognition (ACR). We focus on extending the commonly used major/minor vocabulary (24 classes) to an extended chord vocabulary of seven chord types with a total of 84 classes. In our experiments, we compare joint and separate classification of the chord type and chord root pitch class using one or two separate models, respectively. We perform a large-scale evaluation using various combinations of training and test sets of different timbre complexity. Our results show that ACR with an extended chord vocabulary achieves high f-scores of 0.97 for isolated chord recordings and 0.66 for mixed contemporary popular music recordings. While the joint ACR modeling leads to the best results for isolated instrument recordings, the separate modeling strategy performs best for complex music recordings. Alongside with this paper, we publish a novel dataset for extended-vocabulary chord recognition which consists of synthetically generated isolated recordings of various musical instruments.

1. INTRODUCTION

Automatic chord recognition (ACR) has been actively researched in the field of Music Information Retrieval (MIR) during the last 20 years. ACR algorithms are an essential part of many music applications such as music transcription systems for automatic lead-sheet generation, music education and learning applications, as well as music similarity and recommendation algorithms. In music practice, chord sequences can be played as different chord voicings (selection and order of chord tones) on a large variety of musical instruments, each with its own unique sound characteristic. Therefore, the biggest challenge in ACR is to extract the predominant harmonic changes in a music signal while being robust against different instrument timbres. Furthermore, tuning deviations of music recordings as well as inherent ambiguities between different chords can complicate the task even more [1]. In general, ACR is approached as a two-step problem. First, the acoustic modeling step deals with the prediction of chord labels from short-term audio signal frames. Secondly, during the temporal modeling step, post-processing algorithms are applied to merge frame-level predictions to longer segment-level chord annotations.

As the first main contribution of this paper, we investigate the under-explored task of recognizing seventh chords as an extension to commonly used major and minor chords. Most previous publications focus on recognizing the 24 possible major and minor chords. In the extended-vocabulary ACR scenario, we investigate 7 different chord types including four seventh chord types and the power-chord, which leads to a total of 84 classes. Throughout this paper, we solely focus on improving the acoustic modeling for ACR and do not apply any temporal modeling algorithms. As a second contribution, we compare joint and separate modeling of the chord root pitch class and the chord type as two possible strategies for ACR which are described in Section 3.2. Finally, we publish a novel dataset alongside with this paper that includes synthetically generated chord sequences of the investigated 7 different chord types played with different chord voicings on various keyboard and guitar instruments. [1]

2. RELATED WORK

Early algorithms for acoustic modeling in ACR use template matching in chromagram representations, which encode the local saliency of different pitch classes in audio signals [1, 2]. Here, musical knowledge about the interval structures in different chord types is used to design chord templates for template matching algorithms. We refer the reader to [3] for a systematic overview over traditional techniques for feature extraction and pattern matching in ACR systems and the importance of pre-processing and post-processing steps.

In contrast, fully data-driven approaches based on deep neural network architectures have been lately shown to outperform hand-crafted feature representations. For instance, Convolutional Neural Networks (CNN) [4], Recurrent Neural Networks (RNN) [5, 6], and Feed-Forward Neural Networks (DNN) [7] are used as the acoustic modeling part. Most CNN-based approaches follow the VGG-style architecture [8] with a sequence of 2D convolutional layers and

[1] The dataset can be accessed at https://www.idmt.fraunhofer.de/en/business_units/m2d/research.html.

max pooling layers for a gradual down-sampling in the time-frequency space. Common time-frequency representations such as Short-time Fourier Transform (STFT) [9], Constant-Q transform (CQT) [4] or its multi-channel extension Harmonic CQT [10] are used as two-dimensional input to the CNN models.

As we focus on the acoustic modeling in ACR algorithms, we only briefly review temporal modeling techniques here. The first approaches for temporal modeling in ACR systems have used techniques from automatic speech recognition such as Hidden Markov models (HMMs) [11, 12]. Recently, Korzeniowski & Widmer use RNN-based chord language and duration models as post-processing after a CNN-based acoustic model [13]. Wu & Li combine a bi-directional Long Short-Term Memory (LSTM) network for sequence modeling and Conditional Random Field (CRF) to infer the final chord label sequence [10].

In real-life music recordings, the occurrence of different chord types is heavily imbalanced. While major and minor chords make up the bulk of annotated chords in available chord recognition datasets, other chord types such as seventh chords are heavily underrepresented. Hence, it becomes hard to train ACR systems to detect such chord types. If ACR algorithms should for instance be used to analyze jazz-related music styles, it becomes mandatory to extend the chord vocabulary by seventh chords. Only a few publications such as [10, 14–16] focus on extended-vocabulary chord recognition and go beyond the common 24 class major/minor chord vocabulary. In order to facilitate training models for the extended-vocabulary ACR, we created and published a novel dataset for large-scale chord recognition which will be detailed in Section 4.2.

3. SYSTEM OVERVIEW

3.1 Input Features

Audio signals with a sample rate of 44.1 kHz are converted into Short-time Fourier Transform (STFT) magnitude spectrograms using a blocksize of 8192 (186 ms), a hopsize of 4410 (100 ms), and a Hann window. The phase is discarded. Using a triangular filterbank, the spectrogram is mapped to a logarithmically-spaced frequency axis with 133 frequency bins and a resolution of 24 bins per octave as in [9]. Logarithmic magnitude compression is used to increase the invariance to dynamic fluctuations in the music signal. Spectral patches are extracted with a blocksize of 15 (1500 ms) and a hopsize of 4 (400 ms) and fed as two-dimensional input to the CNN model.

3.2 Modeling Strategies & Network Architecture

Figure 1 shows the CNN model architecture, which we adopted from [9]. As shown in Figure 2, we compare two modeling strategies for ACR: In the first strategy (**S1**), we aim to directly classify the chord label and use a single-output model. Depending on the chord vocabulary size, the final dense layer has either 24 units for classifying major & minor chords or 84 units for classifying all 7 chord types listed in Table 1 given all possible 12 chord root pitch classes. In the second strategy (**S2**), we predict the chord root pitch class (12 classes) and chord type (2 or 7 classes) using two separate models. In both scenarios, the final dense layers have a softmax activation function and after

Abbreviation	Chord Type (# Chord Tones)
5	"Power-chord" (2)
maj	Major chord (3)
min	Minor chord (3)
maj7	Major-seventh chord (4)
min7	Minor-seventh chord (4)
dom7	Dominant-seventh chord (4)
m7b5	Half-diminished seventh chord (4)

Table 1. Investigated chord types with the corresponding number of chord tones.

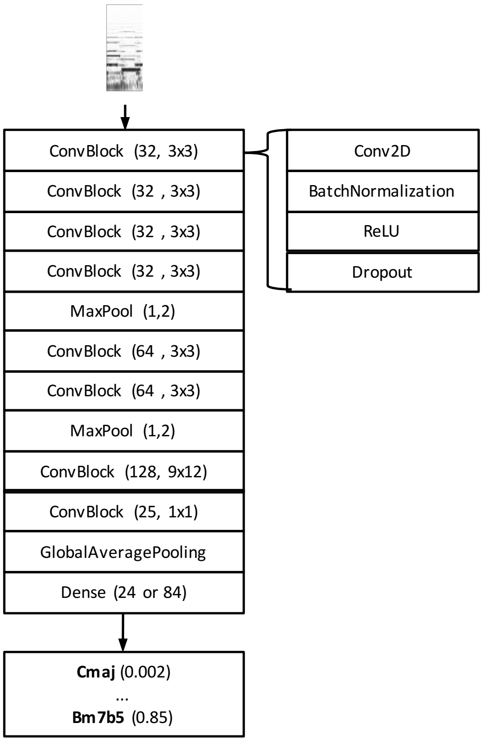

Figure 1. Architecture of the applied CNN. Number of filters and the kernel size are given in brackets for each ConvBlock. The softmax activation function is used in the final dense layer.

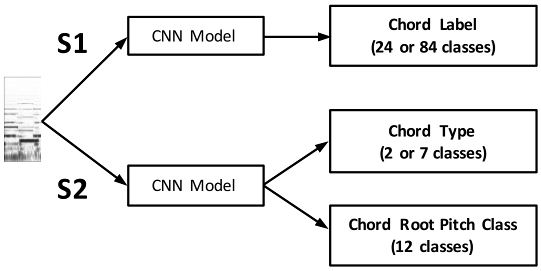

Figure 2. Illustration of two modeling strategies **S1** and **S2** for joint and separate chord type & root pitch class estimation.

Dataset		# Files	Duration (h)	# Chord Segments
Bs	Beatles	1152	53.1	86868
Qn	Queen	180	11.2	20610
RW	Robbie Williams	234	19.1	25569
RWC	RWC	900	61.0	110331
Os	Osmalsky	7200	3.8	7200
Combi7	Combined Dataset	1863	112.2	193194
ISGuitar	IDMT_SMT_GUITAR	48	32.1	684
ISChords	IDMT_SMT_CHORDS	16	4.1	7398
ISInhouse	IDMT Inhouse Dataset	111	4.9	9159

Table 2. Overview of all chord recognition datasets with the respective number of audio files, the total duration in hours, as well as the number of chord segments.

each convolutional layer batch normalization [17] and a rectified linear unit (ReLU) activation function is applied. During training, we use the categorical cross-entropy loss, 500 training epochs with early stopping, the Adam optimizer [18] with a learning rate of 0.003, and a batch size of 256. The input features were normalized to zero-mean and unit-variance for the whole training set. The normalization values were later applied to the test data. All experiments were conducted using the Keras framework with Tensorflow as backend. [2]

4. DATASETS

4.1 Existing Datasets

The datasets used in this paper are summarized in Table 2. In addition to the total number of files, Table 2 provides the total dataset duration and total number of chord segments. In order to enlarge the dataset, we use pitch-shifting with total shifts of up to 4 semitones upwards and downwards as data augmentation technique. Hence, each original file results in 9 augmented files including the original recording. The datasets Beatles (Bs) [19], Queen (Qn) [19], Robbie Williams (RW) [20], RWC (100 songs from the RWC Popular Music Database [21]), and Osmalsky (Os) [22] have been used in the chord recognition literature previously. While the first four datasets include mixed music recordings with multiple instruments, the Os as well as the ISGuitar dataset (excerpts from the IDMT_SMT_GUITAR database published in [23]) consist of isolated recordings of different instruments playing chords. We created and published a novel dataset for chord recognition research (IDMT_SMT_CHORDS, abbreviated as ISChords in this paper), which will be detailed in the following section 4.2. The ISInhouse dataset is an inhouse dataset covering various pop and rock music recordings, which cannot be published due to copyright constraints. In order to evaluate our model on music mixtures for the task of extended-vocabulary ACR, we aggregated an additional dataset (Combi7) using files which include seventh chord annotations from the datasets Bs, Qn, RWC, RW, and Os.

Dataset	maj	min	maj7	min7	5	dom7	m7b5
Bs	67.95	20.49	2.17	3.00	0.04	6.12	0.22
Qn	63.81	22.52	1.28	4.47	1.28	6.56	0.09
RW	69.64	28.30	0.36	0.50	0.89	0.32	-
RWC	48.09	26.57	5.94	13.25	-	5.87	0.28
Os	60.00	40.00	-	-	-	-	-
Combi7	53.24	25.20	4.65	9.75	0.19	6.71	0.27
ISGuitar	67.89	16.51	4.59	3.67	-	5.50	1.83
ISChords	17.11	17.11	14.31	14.31	8.55	14.31	14.31
ISInhouse	62.85	37.15	-	-	-	-	-

Table 3. Chord type distribution per dataset in percent (%).

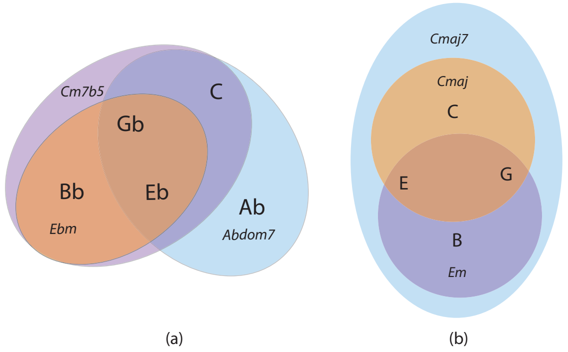

Figure 3. Illustration of ambiguities between chords due to shared chord tones between the chord types m7♭5, min, and dom7 (a), and maj7, maj, and min (b). Figure inspired by [1].

4.2 Synthetic Dataset for Extended-Vocabulary Chord Recognition

Currently used chord recognition datasets are only partially suitable for training and evaluation on seventh chord types. Therefore, we created and published the novel IDMT_SMT_CHORDS dataset [3]. We initially created two MIDI files which cover all seven chord types listed in Table 1. Here we focused on chord voicings, which are commonly used on keyboard instruments and guitars. The piano MIDI file includes all chord types in all possible root note positions and inversions. The guitar MIDI file is based on barré chord voicings with the root note located on the low E, A, and D strings. We used several software instruments from Ableton Live [4] and Garage Band [5] to synthesize these MIDI files with various instruments such as piano, synthesizer pad, as well as acoustic and electric guitar.

5. EVALUATION

In the experiment described in this section, we focus on two types of ACR challenges: First, as discussed in [1], the assignment of a chord label is often ambiguous as different chord types partly share chord tones. Figure 3 illustrates these ambiguities for two chord types which share multiple chord tones. For instance, as shown on the left side, a half-diminished seventh chord (e. g., Cm7♭5) can potentially be confused with different pitch classes like the

[2] Keras: keras.io, Tensorflow: www.tensorflow.org

[3] The download link for the audio and MIDI files will be published in the camera-ready version of the paper.

[4] https://www.ableton.com/en/live/

[5] https://www.apple.com/mac/garageband/

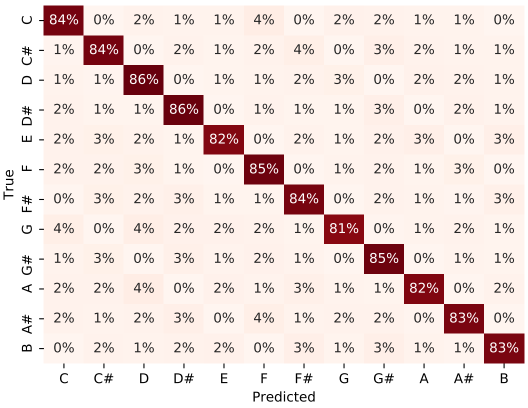

Figure 4. Confusion matrix for chord root pitch class classification on isolated chord recordings (ISChords, experiment **E3**, strategy **S2**).

minor chord built upon its minor third (E♭m) or with the dominant seventh chord built by introducing (A♭) as a root note (A♭dom7).

Secondly, the datasets introduced in Section 4 have different acoustic characteristics. While some of the songs in the Bs and Qn datasets were recorded in the 1970s, other datasets such as RW and ISInhouse contain contemporary popular music recordings with a modern sound. Also, the datasets are of different timbre complexity ranging from simple isolated chords to complex audio mixtures. It was observed in related MIR tasks such as music transcription [24] that data-driven models trained for transcribing isolated notes do not generalize well to more complex acoustic mixtures. Here, we aim to investigate whether such findings can be replicated for ACR.

Table 4 summarizes 11 experiments, which are designed to analyze the chord type ambiguity on isolated chord recordings (**E1** - **E3**, see Section 5.1), the generalization of ACR models to mixture recordings (**E4** - **E6**, see Section 5.2), as well as two real-life ACR application scenarios (**E7** - **E11**, see Section 5.3). In addition, we tested the state-of-the-art ACR algorithm proposed in [9] as our reference system (**REF**) for the major/minor chord vocabulary (24 classes). The implementation from the madmom [25] python library was used and its performance is documented in the last column of Table 4.

In all experiments, audio recordings are split into training and test set on a dataset-level or on a file-level. When a dataset is used for training and test we perform a two-fold random cross-validation. We use the weighted average class f-score throughout this paper as evaluation measure. The f-scores F_{24} and F_{84} are used to indicate if the evaluation was performed on 24 chord classes (major/minor vocabulary) or 84 classes (extended-vocabulary ACR). The "no chord" class is neglected in all experiments. In the following subsections, three groups of experiments will be detailed whose results are summarized in Table 4.

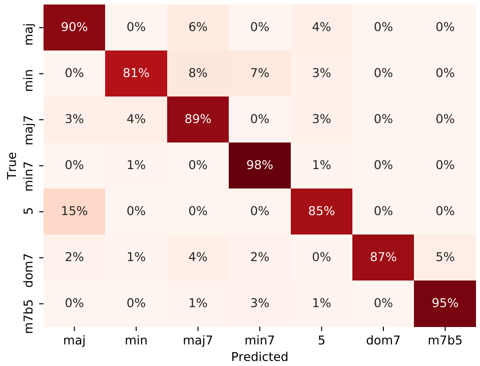

Figure 5. Confusion matrix for 7 chord types in extended-vocabulary ACR on isolated chord recordings (ISChords, experiment **E3**, strategy **S2**).

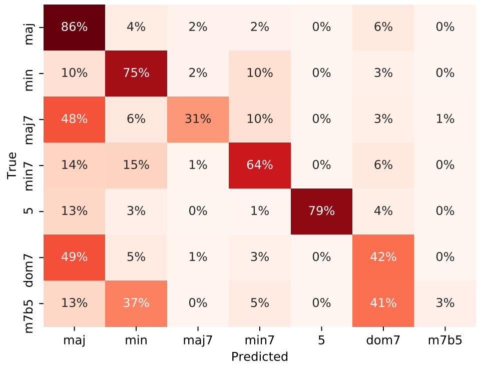

Figure 6. Confusion matrix for 7 chord types in extended-vocabulary ACR on mixed chord recordings (Combi7 + ISChords, experiment **E6**, strategy **S2**).

5.1 Chord Type Ambiguity on Isolated Chord Recordings

In experiments **E1**, **E2**, and **E3** (first section of Table 4), we train and evaluate ACR models on isolated chord recordings (ISChords) to study the effect of chord tone ambiguity in extended-vocabulary ACR. As explained in Section 4.2, the contained chords are based on two systematically generated MIDI files with chord voicings from keyboard and non-keyboard instruments. In our experiments, we evaluate the influence of the chord voicing types as well as of the modeling approach (compare Section 3.2).

For the major/minor chord vocabulary (24 classes), we obtain high f-scores F_{24} between 0.81 and 0.99 using the strategy **S1**. In the two experiments **E1** & **E2**, we perform a chord voicing "cross-test" by exclusively assigning piano chord voicings to the training set and test on non-piano chord voicings and vice versa. Intuitively, we observe lower f-scores (compared to **E3**) since the models are confronted with a different timbre (instrument) and previously unseen chord voicings at test time. Contrary to the 24 classes major/minor scenario, we observe that for the 84 classes scenario (extended-vocabulary ACR),

#	Training Set	Test Set	Strategy S1		Strategy S2		Reference System (REF)
			F_{24}	F_{84}	F_{24}	F_{84}	F_{24}
Chord Type Ambiguity on Isolated Chord Recordings (Section 5.1)							
E1	ISChords (non-guitar)	ISChords (guitar)	**0.92**	0.58	0.90	**0.76**	0.74
E2	ISChords (guitar)	ISChords (non-guitar)	**0.81**	0.49	0.54	**0.56**	0.71
E3	ISChords	ISChords	**0.99**	**0.97**	0.90	0.82	0.75
Generalization of ACR Models towards Complex Recordings (Section 5.2)							
E4	ISChords	Combi7	**0.40**	**0.36**	0.18	0.28	0.83
E5	Combi7	Combi7	0.83	0.63	**0.84**	**0.64**	0.83
E6	ISChords + Combi7	ISChords + Combi7	0.84	0.65	**0.84**	**0.66**	0.81
Real-Life ACR Application Scenarios (Section 5.3)							
E7	ISChords	ISInhouse	0.56	-	0.27	-	**0.76**
E8	ISChords	ISGuitar	0.90	-	0.70	-	**0.91**
E9	Bs + Qn + RW + RWC + Os +ISChords	ISInhouse	0.71	-	0.74	-	**0.76**
E10	Bs + Qn + RW + RWC + Os + +ISChords	ISGuitar	0.90	-	**0.91**	-	**0.91**
E11	Bs + Qn + RW + RWC + Os +ISChords	Bs + Qn + RW + RWC + Os +ISChords	0.81	-	**0.84**	-	0.78

Table 4. This table lists all ACR experiments grouped into three sections described in Section 5.1, Section 5.2, and Section 5.3. For each experiment, the second and third column introduce the applied training set and test set. For both modeling strategies **S1** and **S2** introduced in Section 3.2, f-scores F_{24} and F_{84} are provided for the 24 classes major/minor chord vocabulary and the 84 classes extended-vocabulary with the 7 chord types as listed in Table 1. For each experiment, the best scores for each of the vocabulary are highlighted using bold font. The last column shows the f-score using the reference system (**REF**) on the test set.

strategy **S2** clearly outperforms **S1**. We assume that the network capacity is large enough to learn distinct spectral patterns for classifying among 24 chord labels. For the extended-vocabulary scenario however, the amount of 84 classes is presumably too high to be learnt by one model using strategy **S1**. Instead, splitting the classification task into two easier sub-tasks (with not more than 12 classes each) using strategy **S2** seems slightly beneficial here. Interestingly, in experiment **E3**, where all chord voicings are mixed, strategy **S1** outperforms strategy **S2** in both the 24 and 84 classes scenarios. When testing with state of the art model (**REF**) in **E3** we see that **REF** does not perform as well since it is likely trained on complex audio mixtures.

Figure 4 shows the confusion matrix for the classification of the chord root pitch class for the 84 class scenario for experiment **E3**. It can be observed that the model shows a good performance for all classes between 82 % and 86 %. Similarly, as can be seen in Figure 5, the model easily learns to distinguish between different chord shapes for isolated chord recordings (ISChords dataset). However, Figure 6 shows the more complicated test case of mixed audio recordings (Combi7 + ISChords datasets). The most prominent misclassifications between the maj7 towards the maj, the dom7 towards the maj, as well as the m7♭5 towards the min and the dom7 all confirm the chord tone ambiguities discussed in Section 5.

5.2 Generalization of ACR Models towards Complex Recordings

In experiments **E4** to **E6** (second section of Table 4), we investigate (similar to [24]) whether and to what extent ACR models trained on isolated instrument recordings generalize towards complex music recordings in the Bs, Qn, and RWC datasets. Also, we test whether adding the proposed ISChords dataset can help to improve the performance on extended-vocabulary ACR. As expected, a poor f-score of $F_{24} = 0.4$ in **E4** shows that the investigated CNN-based ACR model does not generalize well from a simple training scenarios (ISChords) towards a complex test scenario (Combi7). The clearly higher f-scores of $F_{24} = 0.84$ and $F_{84} = 0.66$ show that this kind of data-driven classification models need to be trained on data of similar timbre complexity as in the test scenario. We only observe a small improvement of 0.02 (from **E5** to **E6**) for the 84 classes scenario in f-score when training with both datasets (**E6**). The reference algorithm **REF** performs similar to **S2** in **E5** and slightly worse than **S1** and **S2** with a difference of 0.03 in **E6**.

5.3 Real-Life ACR Application Scenarios

In the experiments **E7** to **E11** (third section of Table 4), we address realistic requirements for ACR systems to be deployed in real-life applications. In a music education scenario, musical instruments usually can be directly recorded and analyzed without background sounds. Therefore, we test the chord recognition performance on isolated poly-

phonic electric guitar recordings (ISGuitar), which include both chords and arpeggios. In a music annotation scenario, we evaluate ACR models on a set of 111 contemporary pop and rock music recordings of various instrumentations
(ISInhouse). Similarly to **E4**, we can observe in experiment **E7** that ACR models trained only on isolated chord recordings do not perform well on complex mixtures (ISInhouse). However, such models show a good performance (**S1**, $F_{24} = 0.9$, **S2**, $F_{24} = 0.7$) when being applied to isolated guitar recordings (**E8**). In both test cases, the performance can be clearly improved by adding more datasets to the training set, which reflect a larger variety of music recordings (compare experiments **E9** and **E10**). In both experiments **E9** and **E10**, the reference algorithm **REF** performs almost similar except for **E11** where **S2** achieves a slightly better f-score.

6. CONCLUSIONS

In this paper, we used a state-of-the-art Deep Convolutional Neural Network for ACR. In addition to publishing a novel dataset of isolated chord recordings, we propose an alternative modeling strategy using two models for the separate classification of the chord type and the chord root pitch class. In our experiments, we first evaluate this strategy for the controlled test case of isolated instrument recordings. Most of the chord type misclassifications are due to shared chord tones. The results indicate that ACR even with extended-vocabulary is feasible (f-scores above 0.9), but the performance depends on whether the chord voicings and instrument timbre used in the test set have been learnt by the model before.

In a second set of experiments, we were able to replicate the finding from automatic music transcription that data-driven ACR models need to be trained on data of the same complexity as the expected test data. Models trained on isolated instrument recordings performed poorly on mixed audio data. Finally, we evaluated the CNN model on two separate datasets, which acted as a proxy for deploying an ACR model into real-life production systems for the two use cases music education and music annotation. Here, we achieved high f-scores of 0.91 for isolated guitar recordings and 0.74 for mixed contemporary popular music recordings showing the usefulness for real-life MIR applications.

Acknowledgments

This work has been supported by the German Research Foundation (AB 675/2-1, BR 1333/20-1).

7. REFERENCES

[1] M. Müller, *Fundamentals of Music Processing*. Springer Verlag, 2015.

[2] T. Fujishima, "Real-time chord recognition of musical sound: A system using common lisp music," in *Proceedings of the 1999 International Computer Music Conference (ICMC)*, Beijing, China, 1999, pp. 464–467.

[3] T. Cho and J. P. Bello, "On the relative importance of individual components of chord recognition systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 477–492, 2014.

[4] E. J. Humphrey and J. P. Bello, "Rethinking automatic chord recognition with convolutional neural networks," in *Proceedings of the 11th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, USA, 2012, pp. 357–362.

[5] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Audio chord recognition with recurrent neural networks," in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, Curitiba, Brazil, 2013, pp. 335–340.

[6] S. Sigtia, N. Boulanger-Lewandowski, and S. Dixon, "Audio chord recognition with a hybrid recurrent neural network," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, 2015, pp. 127–133.

[7] X. Zhou and A. Lerch, "Chord detection using deep learning," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, 2015, pp. 52–58.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[9] F. Korzeniowski and G. Widmer, "A fully convolutional deep auditory model for musical chord recognition," in *Proceedings of the 26th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Salerno, Italy, 2016, pp. 1–6.

[10] Y. Wu and W. Li, "Automatic audio chord recognition with midi-trained deep feature and blstm-crf sequence decoding model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 355–366, 2019.

[11] A. Sheh and D. P. Ellis, "Chord segmentation and recognition using em-trained hidden markov models," in *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*, Baltimore, Maryland, USA, 2003.

[12] J. Deng and Y.-K. Kwok, "Large vocabulary automatic chord estimation using deep neural nets: Design framework, system variations and limitations," *arXiv preprint arXiv:1709.07153*, 2017.

[13] F. Korzeniowski and G. Widmer, "Improved chord recognition by combining duration and harmonic language models," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 10–17.

[14] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, 2010, pp. 135–140.

[15] J. Deng and Y. Kwok, "Automatic chord estimation on seventhsbass chord vocabulary using deep neural network," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 261–265.

[16] S. Gasser and F. Strasser, "Mirex 2018: Multi objective chord estimation," 2018, (last accessed 28.03.2019). [Online]. Available: https://www.music-ir.org/mirex/abstracts/2018/SG1.pdf

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[19] "Isophonics dataset reference annotations," (last accessed 24.01.2019). [Online]. Available: http://isophonics.net/datasets

[20] B. Di Giorgi, M. Zanoni, A. Sarti, and S. Tubaro, "Automatic chord recognition based on the probabilistic modeling of diatonic modal harmony," in *Proceedings of the 8th International Workshop on Multidimensional Systems (nDS)*, Erlangen, Germany, 2013, pp. 1–6.

[21] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002, pp. 287–288.

[22] J. Osmalsky, V. D. M. Embrechts, Jean-Jacques, and S. Pierard, "Neural networks for musical chords recognition," *Journées d'informatique musicale*, pp. 39–42, 2012.

[23] "IDMT SMT GUITAR dataset," (last accessed 24.01.2019). [Online]. Available: https://www.idmt.fraunhofer.de/en/business_units/m2d/smt/guitar.html

[24] J. Abeßer, S. Balke, and M. Müller, "Improving bass saliency estimation using label propagation and transfer learning," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 306–312.

[25] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "madmom: a new Python Audio and Music Signal Processing Library," in *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, The Netherlands, 2016, pp. 1174–1178.