Towards Interpreting and Improving the Latent Space for Musical Chord Recognition

Christon R. Nadar **Michael Taenzer** **Jakob Abeßer**
Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany
christon-ragavan.nadar@idmt.fraunhofer.de

ABSTRACT

Automatic chord recognition (ACR) naturally faces musical ambiguities between chord classes. These can be responsible for many misclassifications, especially in large chord vocabularies. In this paper, we propose a metric learning approach utilizing a triplet loss for the task of ACR in order to reduce chord ambiguities. In particular, we investigate how metric learning with different triplet sampling strategies re-aligns the distances between different chord classes in the latent space. Our main finding is that metric learning significantly improves the ACR performance for two taxonomies with five and nine chord classes.

1. INTRODUCTION

Harmony analysis is a key ingredient of many music information retrieval (MIR) tasks such as music transcription, music identification, as well as music similarity and recommendation. As an important subtask, automatic chord recognition (ACR) aims to identify and transcribe a musical chords from audio recording. While this task has been researched for over two decades, most ACR algorithms so far focused on a small vocabulary of 24 chords covering all major and minor chords. However, such simplification does not match the complexity of different chord qualities used in music genres such as jazz or popular music. Therefore, the focus of current research has changed towards large-vocabulary chord transcription, which faces several challenges [1].

First, by considering additional (extended) chords, ambiguities caused by shared chord tones become more evident as some basic chords are subparts of higher degree chords. An example for such an ambiguity is that only three fully diminished chords (`dim7`) with unique chord tones exist. Here, all four chord-tones are three semitones apart from each other. From the perspective of an ACR system, `Cdim7` and `E♭dim7` consist of the same chord tones `C`, `E♭`, `G♭`, and `A`. Chord inversions, i. e., different cyclic permutations of chord tones, can cause additional ambiguities as they share the same chord tones, but have a different interval structure. One example is the pair of chords `Am7/C`, which is the chord interpretation from the root note `A`, and `C6`, which is the chord interpretation from the root note `C`. Another challenge for ACR algorithms is the influence of various instrument timbres and note loudness on the perception of chord qualities. For instance, a predominant bassline may affect which pitch class is perceived as chord root.

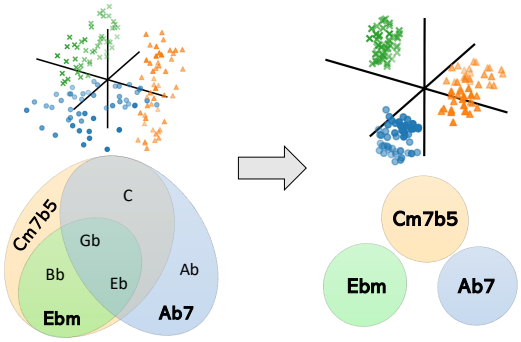


Figure 1: Illustration of chord class separability in the latent space

Finally, the quality of annotations directly affects the performance of data driven methods such as deep neural networks (DNN). The inherent subjectivity of human chord annotations can introduce label noise, another source of bias and ambiguity [1]. Furthermore, ACR labels commonly lack information concerning the octave position of a chord, its note dynamics, as well as its inversion. If simpler chord taxonomies are targeted, complex chord qualities are often simplified and re-mapped to simpler classes, e.g. from `maj13` to `maj7` and eventually to `maj`.

The majority of ACR research nowadays relies on deep learning algorithms, which are trained in a data-driven fashion. However, only little research was conducted to better understand the learnt feature representations. Our main research objective is to interpret typical errors and chord confusions of state-of-the-art large vocabulary ACR algorithms in order to allow for a systematic improvement.

This paper offers two main contributions. First, we introduce several metrics to better understand how different chord qualities are distributed in learnt latent spaces of deep learning based ACR algorithms. In particular, we use a variance ratio inspired by Linear Discriminant Analysis (LDA) and further introduce the "pair-wise distance matrix (PDM)" to quantify the separability between different chord classes. Second, we study a metric learning approach based on the triplet-loss to improve the ACR performance by improving the chord separability in the latent space. Figure 1 is a toy example to illustrate the proposed method to interpret and improve chord class separability.

2. RELATED WORK

Traditional ACR approaches rely on chroma-based audio features, which encode the salience of different pitch classes throughout the audio signal. In earlier works, pattern matching techniques based on pre-defined chord templates were used to determine the chord label [2]. Another common approach combines an acoustic model that focuses on the feature learning and recognition of the chord in a given frame and a temporal model that models the temporal dependencies associated with a given chord frame to estimate harmonic progression [3, 4, 5]. In this work, we solely focus on the acoustic modeling part of ACR systems.

Recent data-driven ACR methods using DNNs have consistently outperformed traditional methods based on handcrafted feature representations. For instance, convolutional neural networks (CNN) [6, 7], recurrent neural networks (RNN) [8, 9, 10], and feed forward neural networks [11] have been applied for acoustic modeling. Widely used signal representations for these purposes are the constant-Q transform (CQT) [8, 9, 10, 11] and magnitude log-frequency spectrogram (MLFS) [5, 12]. To capture the harmonic series of the chord-tones in MLFS representations, spectral weighting on the spectrogram was considered [13, 14] as a pre-processing step. In some cases, the Harmonic CQT (HCQT) [15, 16] and raw waveform as input feature [17] also found application. We recommend [1] for an in-depth review of existing ACR methods.

Musical content can be described by multiple properties such as chords, melody, rhythm, genre, etc. For popular music, these musical dimensions are typically not independent, but in fact highly interdependent. To exploit the correlation of this information, several efforts towards multitask learning approaches have been made. One example of simultaneous estimation for chords and other musical properties such as beat, downbeat, bass pitch class and key is demonstrated in [18]. To exploit the onset information from downbeats, joint estimation of chord and downbeat was also considered [19, 20].

Despite a wide range of methods, earlier ACR algorithms were restricted to a small vocabulary of 24 major and minor chords. Several recent works have shown interest towards a larger chord vocabulary [12, 21, 22, 23, 24]. Increasing the vocabulary comes with an increase of chord confusions, as mentioned before.

In recent works, methods employing Deep Metric Learning (DML) aim to learn latent spaces, where the proximity of data instances better reflect their semantic similarity. The triplet loss function is most often used in this regard [25, 26, 27]. For our work, we use a triplet model to improve the separability of the chords, thereby reducing chord confusion and improving the overall ACR system.

3. PROPOSED METHOD

3.1 Input Features & Targets

The choice of a suitable signal representation is crucial for DNN based classification methods. During the acoustic modeling step in ACR methods, the neural network learns a mapping $f : \mathbb{R}^{N_T \times N_F} \rightarrow \mathbb{R}^{N_T}$ from a two-dimensional input feature X to frame-level targets y with N_T denoting the number of time frames and N_F denoting the number of frequency bins. All audio signals are processed at a sampling rate of 44.1 kHz and converted to mono. The input feature X is derived by first computing the Short-Time Fourier Transform (STFT) with an FFT size of 8192 samples (186 ms) and a hop size of 4410 samples (100 ms). Then, the magnitude spectrogram is mapped to a logarithmically spaced frequency axis using a triangular filterbank resulting in 133 frequency bins with a resolution of 24 bins per octave. Finally, equal-sized spectral patches of 15 frames (1.5 s) duration are extracted with 50% overlap and the corresponding chord class targets are generated.

We do not simplify or remap existing chord annotations, but rather select annotated frames for two different class taxonomies, which we consider to study further potential chord confusions caused by extended chord vocabularies. First, since extended tetrads often have hidden triads [12], we experiment with the five chord classes `maj`, `min`, `maj7`, `min7`, and `hdim7`. The chord `hdim7` contains both `maj` and `min` triads. Second, a larger vocabulary with nine chord classes is considered, additionally consisting of the chords `7`, `dim`, `dim7`, and `5` ("power chord"). We exclude the "no chord" class in this work.

3.2 Neural Network Architectures

We employ two neural networks. First, a CNN model adopted from [6, 12] is used as a baseline system, denoted as `B`. Second, to improve upon the chord confusion and to increase the separability of chord classes, a triplet model is considered which is explained in Section 3.3. This model is denoted as `T`.

To keep both models comparable, the core CNN model for feature learning was kept similar. It comprises four convolutional blocks, each incorporating four blocks of stacked convolutional layers (ConvLayers). The first block contains four ConvLayers, each with 32 filters and a kernel shape of 3×3. The second block contains two ConvLayers with 64 filters of size 3×3 each. The third block contains one ConvLayer with 128 filters of size 12×9 in order to aggregate harmonic information for the classification part of the network. The final block is used to reduce the feature dimension with one ConvLayer with 25 filters of size 1×1. Each block is followed by a max-pooling layer of size 1×2 and a dropout layer with a dropout ratio of 0.5 to avoid overfitting. Finally, each ConvLayer is followed by batch normalization and the ReLu activation function. To assure comparability with the triplet embeddings, we added an additional dense (embedding) layer with 20 units per chord class. We chose the dimension 20 empirically, such that in the case of 5 chord classes, the dimension of the dense layer would be 100, whereas for 9 chord classes, it is set to 180. Finally, L_2 normalization is applied to the embedding layer.

The baseline model was trained in a supervised fashion using the Adam optimizer with a learning rate of 10^{-4}, a batch size of 128, and the categorical cross entropy loss function.

3.3 Metric Learning and Sampling Strategies

Recently, DML methods have been successfully applied to improve the class separability in latent representation

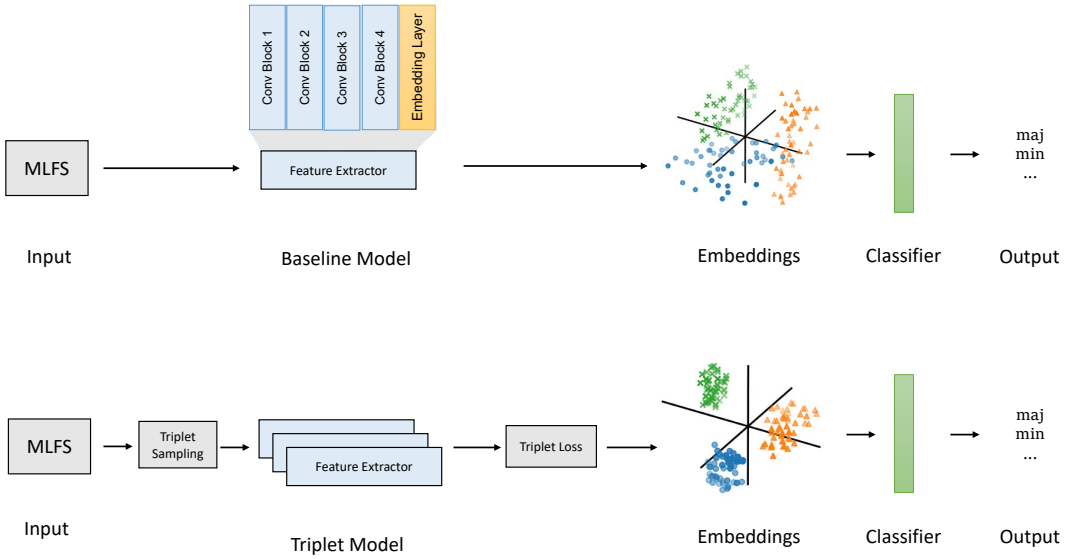


Figure 2: Comparison between a conventional supervised learning based ACR approach (top) and a metric learning based ACR approach (bottom), where the latent representation (embeddings) is learnt prior to the final classification layer.

spaces[26]. Therefore in this work, we follow a triplet-based learning approach to train a core network, which extracts embeddings that are processed by the classifier, as shown in Figure 2. The triplet loss function enforces the model to learn a latent space representation based on data triplets, i.e., an anchor, a positive, and a negative, in order to minimize the distance between the anchor and positive, while simultaneously maximizing the distance between the anchor and negative. The dimension of the embedding layer is chosen similar to the baseline model (see Section 3.2). After the metric learning stage, the final classification layer is trained with the parameters of the embedding extractor layers. We compare two sampling methods to generate data triplets. First, we randomly sample a negative example for a given anchor such that it corresponds to a different chord class (denoted as `random`). Second, we found that most of the seventh chord classes are confused for their respective major and minor chords. Therefore, in the second sampling strategy, denoted as `maj-min`, we randomly sample negatives from the respective major and minor chord of a given anchor. For example, a `maj` chord will be sampled as negative for all `maj7`. In the cases of `hdim7` and `5` chords, both `maj` and `min` can be sampled at a probability of 50%.

4. EVALUATION

4.1 Dataset

In our experiments, we use five publicly available ACR datasets listed in Table 1. The datasets Beatles (`Bs`) [28], Queen (`Qn`) [28], Robbie Williams (`RW`) [30], and `RWC` (100 songs from the RWC Popular Music Database [29]) contain multi-instrumental recordings from popular music genres. The IDMT_SMT_Chords (`Ic`) [12] dataset includes single instrument recordings of chords synthesized from MIDI using several digital audio workstations. The `Ic` dataset was built around two types of chord voicings:

Dataset	Files	Duration
Beatles [28]	180	8h 09m
RWC [29]	100	6h 47m
Robbie Williams [30]	26	2h 06m
Queen [28]	20	1h 12m
IDMT_SMT_Chords [12]	16	4h 6m
Total	326	22h 20m

Table 1: Overview of the datasets. Durations are given in hours (h) and minutes (m).

a) played on keyboard instruments (e.g. triads with three chord-tones), and b) replicating 6-string guitar play style (e.g. barré chords, and open position triads with six chord-tones). We performed a 80%-10%-10% split for the train, validation, and test set on file level. To ensure a balanced distribution of the seventh chords, first the files with extended chords were segregated from files with only maj and min chords. Further, another 80%-10%-10% train, validation and test split was performed on both of the segregated sets, which were later combined to their respective train, validation and test sets. We furthermore generate eight pitch shifted versions of each file within a range of +/- 4 semitones as data augmentation.

4.2 Metrics

Throughout this paper, we report frame-level (ACR) performance using the F1-score F. To account for the large class imbalance of chord qualities found in the real-life music recordings, we use the weighted macro F1-score, where each class is averaged individually and weighted by the number of true targets. Furthermore, we use two metrics to characterize the class distribution in the latent space, which are inspired by Linear Discriminant Analysis (LDA). The first metric is the between-class variance σ_{b}, computed based on the Euclidean distances between the

class centroids. As a second metric, the within-class variance σ_{w} is computed per class to estimate the spread of its data points, and finally averaged over all classes. Lastly, a ratio (σ_{r}) is computed:

$$\sigma_{\mathrm{r}} = \frac{\sigma_{\mathrm{b}}}{\sigma_{\mathrm{w}}}. \tag{1}$$

4.3 Improving Class Separability in the Latent Space

Model	Sampling method	# Classes	$\sigma_{\mathrm{b}} \uparrow$	$\sigma_{\mathrm{w}} \downarrow$	$F \uparrow$
B	-	5	0.41	**0.18**	0.70
B	-	9	0.44	**0.08**	0.60
T	random	5	0.44	**0.18**	0.71
T	random	9	0.45	**0.08**	0.63
T	maj-min	5	**1.28**	0.42	**0.73**
T	maj-min	9	**2.49**	0.38	**0.66**

Table 2: Influence of model training and triplet sampling strategy on the between-class variance σ_{b}, the within-class variance σ_{w}, and the F1-score F of the ACR systems.

In this experiment, we investigate the influence of the proposed DML method on the data distribution in the latent space (measured by σ_{b} and σ_{w}) as well as on the ACR performance (measured by F). We compare the baseline model B and the triplet model T as described in Section 3.3. The metrics σ_{b} and σ_{w} for model B were computed based on the embedding layer output for the test dataset. The triplet model T is trained using the two triplet sampling methods described in Section 3.3.

As can be seen in Table 2, the model T improves over model B in F1-score. The maj-min sampling strategy leads to the best ACR performance of $F = 0.73$ and $F = 0.66$ for the 5-class and 9-class chord taxonomies, respectively. This strategy shows a clear increase in σ_{b} compared to the other two configurations, which indicates an improved class separability in the latent space. This comes at the cost of an increase of the within-class variance σ_{w}, however, to a smaller extent.

4.4 Influence of the Network Depth

		5 chord classes				9 chord classes			
Model	N_{D}	$\sigma_{\mathrm{b}} \uparrow$	$\sigma_{\mathrm{w}} \downarrow$	$\sigma_{\mathrm{r}} \uparrow$	$F \uparrow$	$\sigma_{\mathrm{b}} \uparrow$	$\sigma_{\mathrm{w}} \downarrow$	$\sigma_{\mathrm{r}} \uparrow$	$F \uparrow$
B	1	0.41	0.18	2.28	0.70	0.44	0.08	5.38	0.60
T	1	1.28	0.42	**3.06**	**0.73**	2.49	0.38	6.40	**0.66**
T	2	1.60	0.56	2.81	0.73	3.35	0.49	6.73	0.66
T	4	2.0	0.68	2.92	0.73	3.06	0.46	6.63	0.66
T	6	2.30	0.84	2.73	0.73	4.37	0.68	6.37	0.65
T	8	2.37	0.85	2.76	0.73	3.42	0.54	6.33	0.65
T	10	2.34	0.85	2.73	0.73	3.58	0.49	**7.20**	0.65

Table 3: ACR performance for different variants of the triplet-based model T with N_{D} dense layers for both the 5-class and 9-class chord vocabularies.

We have shown in Section 4.3 that model T clearly outperforms model B in ACR performance. In addition to the general training methodology, we aim to investigate in this experiment the influence of the network depth, i. e., the number of dense layers prior to the final classification layers. Here, we consider model T with the maj-min triplet sampling strategy. The learnt weights from the initial part of the network until the embedding layer were reused to train the different network variants with additional dense layers "on top". Each dense layer is followed by a ReLu activation and has the dimension of 20 per chord class as discussed in Section 3.2. Table 3 shows the results of the network T for $N_{\mathrm{D}} \in \{1, 2, 4, 6, 8, 10\}$ additional dense layers.

For the small vocabulary of 5 chord classes, the network has reached the optimum learning capacity for a shallow model ($N_{\mathrm{D}} = 1$) as we cannot see any improvement w.r.t. F for a larger number of layers. Here, an increase in the number of layers goes along with an increasing between-class variance but also an increasing within-class variance. As shown by the σ_{r} values, both effects combined lead to a decreased class separability for larger N_{D}. In the case of the large chord vocabulary of 9 classes, deeper networks show an improved class separability in the latent space, which is supported by the highest value of $\sigma_{\mathrm{r}} = 7.20$ for 10 layers. However, this does not lead to an improved ACR performance.

4.5 Musical Interpretation of Chord Class Confusions

In this experiment, we aim to investigate to what extent the musical similarity between different chord types is captured by their distances in the latent space of the best performing ACR model T. We expect the results to reveal potential weak spots of the learnt representation, which can cause chord confusions.

Therefore, we investigate the pair-wise distances between class centroids in the latent space. In the following, $z_i \in \mathbb{R}^L$ denotes the embedding representation of the i-th example in the L-dimensional latent space. The corresponding chord class is denoted by $y_i \in \mathbb{Z}^{N_c}$ and the number of examples and chord classes is N and N_c, respectively. We compute a *pair-wise distance matrix* (PDM) $P \in \mathbb{R}^{N_c \times N_c}$ as follows: Based on the class centroids

$$\tilde{z}_c = \frac{1}{N^c} \sum_{i=1}^{N} \begin{cases} z_i & \text{if } y_i \equiv c, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

with $c \in [1 : N_c]$, the pair-wise Euclidean distances between chord classes are computed as

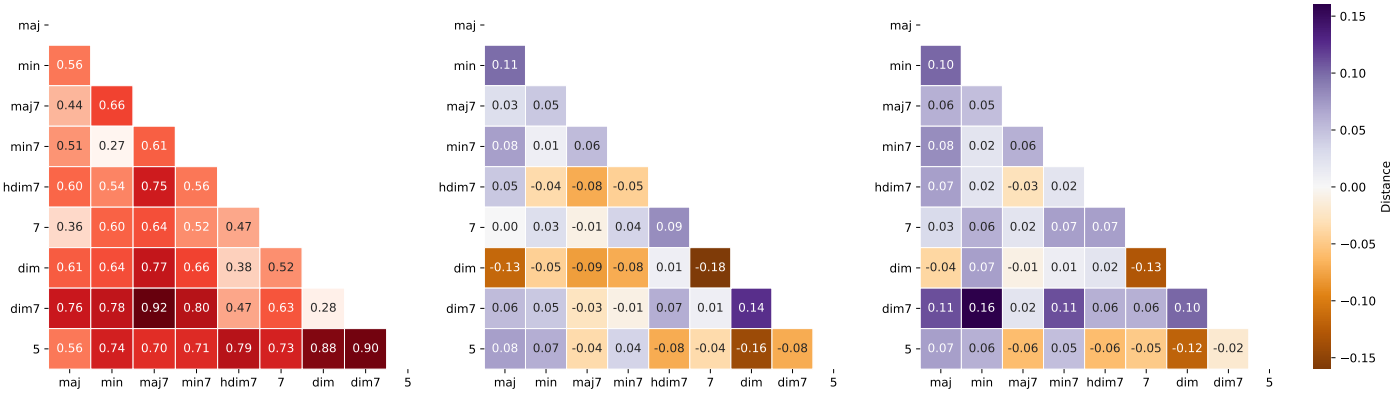
$$P_{i,j} = \|\tilde{z}_i - \tilde{z}_j\| \text{ for } i, j \in [1 : N_c]. \tag{3}$$

To better understand which particular chord misclassifications were improved using the triplet-based model, we compute the relative PDM between both networks as

$$\Delta P_{\mathrm{T,B}} = P_{\mathrm{T}} - P_{\mathrm{B}}. \tag{4}$$

Figure 3a illustrates the pairwise distance matrix P for the baseline model B. Furthermore, the relative PDMs $\Delta P_{\mathrm{T,B}}$ between the triplet-based models are shown for both the random (3b) and maj-min (3c) triplet sampling strategies and the baseline model. Due to their symmetric nature, only the lower triangles of P and ΔP are shown.

From these visualizations, we make the following observations. First, lower values in P_{B} reveal pairs of similar chord classes such as min7 and min, 7 and maj, as well as dim and hdim7, which are mostly plausible according to shared chord tones and interval structures. When looking into the improvements of the triplet-based networks



(a) (b) (c)

Figure 3: Pairwise Distance Matrix (PDM) P for the baseline model `B` (a). Relative PDMs $\Delta P_{\mathrm{T,B}}$ between the triplet-based model `T` with `random` (b) or `maj-min` (c) triplet sampling strategy and the baseline model `B`.

over the baseline model, we can see that many chord pairs such as `min-maj`, `7-hdim7`, as well as `5-maj` show a larger distance and hence a better discriminability. The biggest improvement can be observed for the `dim7` chord with reduced confusions to the `dim` chord (`random` sampling), as well as to the `maj`, `min`, `min7`, and `dim` chords (`maj-min` sampling).

5. CONCLUSIONS

In this paper we used a triplet based deep metric learning approach for the task of Automatic Chord Recognition (ACR) to address the problem of chord confusion. We presented two triplet sampling strategies for reducing the chord confusion in the latent space. To interpret, understand and evaluate the chord overlap in the latent space, we presented two approaches. First, we used an LDA based within-class variance and between-class variance. Furthermore, a Pair-Wise Distance Matrix (PDM) is computed to evaluate the class separation in the latent space. Our results show that using triplets there is a clear improvement over an existing ACR baseline model.

6. ACKNOWLEDGEMENTS

This work has been supported by the German Research Foundation (AB 675/2-2).

7. REFERENCES

[1] J. Pauwels, K. O'Hanlon, E. Gómez, and M. B. Sandler, "20 Years of Automatic Chord Recognition from Audio," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, 2019, pp. 54–63.

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

[3] 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.

[4] 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, pp. 185–191.

[5] F. Korzeniowski and G. Widmer, "Feature Learning for Chord Recognition: The Deep Chroma Extractor," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR, New York City, United States, August 7-11, 2016*, 2016, pp. 37–43.

[6] 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.

[7] 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.

[8] J. Jiang, K. Chen, W. Li, and G. Xia, "Large-vocabulary Chord Transcription Via Chord Structure Decomposition." in *ISMIR*, 2019, pp. 644–651.

[9] 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, 2018.

[10] 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.

[11] 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.

[12] C.-R. Nadar, J. Abeßer, and S. Grollmisch, "Towards CNN-based acoustic modeling of seventh chords for automatic chord recognition," in *International Conference on Sound and Music Computing. Málaga, Spain*, 2019.

[13] K. Lee, "A system for automatic chord transcription from audio using genre-specific hidden Markov models," in *International Workshop on Adaptive Multimedia Retrieval*. Springer, 2007, pp. 134–146.

[14] J. Morman and L. Rabiner, "A System for the Automatic Segmentation and Classification of Chord Sequences," in *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, ser. AMCMM '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 1–10. [Online]. Available: https://doi.org/10.1145/1178723.1178725

[15] 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.

[16] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep Salience Representations for F0 Estimation in Polyphonic Music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 63–70.

[17] J. Pons, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-end Learning for Music Audio Tagging at Scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. Paris, France: ISMIR, Sep. 2018, pp. 637–644. [Online]. Available: https://doi.org/10.5281/zenodo.1492497

[18] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1280–1289, 2009.

[19] H. Papadopoulos and G. Peeters, "Joint estimation of chords and downbeats from an audio signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 138–152, 2010.

[20] H. Papadopoulos and G. Peeters, "Simultaneous estimation of chord progression and downbeats from an audio file," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 121–124.

[21] G. Byambatsogt, L. Choimaa, and G. Koutaki, "Guitar Chord Sensing and Recognition Using Multi-Task Learning and Physical Data Augmentation with Robotics," *Sensors*, vol. 20, no. 21, p. 6077, 2020.

[22] M. Bortolozzo, R. Schramm, and C. R. Jung, "Improving the Classification of Rare Chords With Unlabeled Data," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3390–3394.

[23] B. McFee and J. P. Bello, "Structured Training for Large-Vocabulary Chord Recognition." in *ISMIR*, 2017, pp. 188–194.

[24] K. M. Kinnaird and B. McFee, "Automatic Hierarchy Expansion for Improved Structure and Chord Evaluation," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, 2021.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[26] J. Lee, N. J. Bryan, J. Salamon, Z. Jin, and J. Nam, "Disentangled multidimensional metric learning for music similarity," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6–10.

[27] N. Turpault, R. Serizel, and E. Vincent, "Semi-supervised triplet loss based learning of ambient audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 760–764.

[28] "Isophonics Dataset Reference Annotations," (last accessed 24.01.2019). [Online]. Available: http://isophonics.net/datasets

[29] 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.

[30] 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.