ANNOTATING JAZZ RECORDINGS USING LEAD SHEET ALIGNMENT WITH DEEP CHROMA FEATURES

Ivan Shanin, Simon Dixon

Centre for Digital Music, Queen Mary University of London, UK

ABSTRACT

The automatic recognition of chords from jazz recordings remains largely an unsolved challenge, due to the broad harmonic vocabulary, the freedom of interpretation in performance, the rich variety of expressive techniques, and the limited availability of accurately labeled training data. In practice, many jazz recordings contain performances of popular compositions that are known as standards. We propose an approach that takes into consideration available prior information about chord changes of popular compositions known as lead sheets. Instead of estimating the exact chord symbol at each time point, we aim to identify the position in the lead sheet, thereby solving the audio-to-score alignment task (with a distinction that the score does not contain any melodic information, and the harmonic annotation is only approximate). This approach also solves the structural segmentation problem, as segment boundaries are available in the lead sheets. To achieve this goal we combine a multi-task convolutional recurrent neural architecture with an alignment algorithm based on a Hidden Markov Model. The proposed approach uses the iRealPro corpus of 1186 lead sheets and is evaluated on a test set of the 220 audio excerpts in the Weimar Jazz Database, showing that it outperforms previously published work.

Index Terms— Music information retrieval, automatic chord estimation, music synchronization

1. INTRODUCTION

An automatic chord estimation algorithm (ACE) [1, 2, 3] provides an informative representation of harmonic context. While previous researchers used signal processing and statistical methods [4, 5], current state-of-the-art approaches use feature representations extracted from short segments of audio as an input for a convolutional recurrent [6] or a transformer [7] neural network to assign a chord class to each frame. Training such a network requires a large dataset, annotated by experts, and the trained model heavily depends on the musical genres selected for the training set [8].

However, ACE approaches show limited performance when applied to jazz recordings [9], since the harmony is much more complex. In pop music, a harmony instrument (usually guitar or keyboard) typically plays the notes of the chord for its full duration. Jazz accompaniment, in contrast, uses extensions (adding extra notes), alterations (modifying the pitch of some chord tones) and substitutions (replacing one chord with another that has a similar harmonic function), and such chords are played sporadically over the chord's nominal duration, interspersed with other chords and tones. As a result, the chord annotation task is quite ambiguous even for a musical expert.

One promising approach to bridge this gap is to use additional information available in a lead sheet (a simplified musical score) and estimate the current position in the musical form using audio to score alignment. This differs from the "forced alignment" used for creating training data [10], where the full sequence for alignment is assumed to be known. In our application (like in [11]), the lead sheet contains the chord sequences for each section, but not the sequencing of the sections, which varies between different performances of the tune. The position in the lead sheet intuitively reflects how music functions: musicians choose notes to play based on their interpretation of the harmony at each moment, while the listener might not even know specific musical terminology, but they both normally understand "where in the song they are" at a given moment.

Lead sheet alignment was previously addressed in [12], where chroma features were used in combination with the Dynamic Time Warping alignment algorithm. The performance of this system was evaluated on a dataset that consisted of several different performances of three jazz compositions. Recently, a convolutional recurrent chord estimator was used to generate lead sheets from the audio [9], using prior information about structure (chorus boundaries). In this work we are developing a reverse approach: using a chord estimator and a lead sheet we aim to detect chorus, section and measure boundaries in the audio. This scenario is especially practical in the automatic annotation of large corpora of jazz recordings, where usually song title is known in advance and the lead sheets for many songs are available in open-source collections.

2. STRUCTURAL SEGMENTATION OF JAZZ RECORDINGS

We follow the terminology proposed by the authors of the Jazz Structure Dataset [13]. Popular jazz compositions (also known as jazz standards) are usually comprised of consecutive sections (known as choruses) that share a common harmonic structure. The first and the last choruses usually contain the main theme, and the rest of the choruses feature improvised solos that emphasize key elements of the same harmonic schema.

Sections within a chorus are labelled by capital letters: for example, AABA indicates a sequence of four sections where sections 1,2 and 4 share the same sequence of chords (as in "I Got Rhythm" by George Gershwin). It is not uncommon that musicians choose to deviate from this form to have better control over the recording duration or for other artistic purposes, but while the sequence of sections within a chorus may sometimes vary, the duration and basic harmonic schema of these sections usually remain unchanged. Apart from that artists sometimes expand the chorus structure by including additional structure elements, such as "intro", "outro",

This research was supported by the AHRC/NEH-funded project "New Directions in Digital Jazz Studies: Music Information Retrieval and AI Support for Jazz Scholarship in Digital Archives" (AHRC grant AH/V009699/1, NEH grant HC-278112-21).

979-8-3503-2372-6/23/$31.00 ©2023 Crown

"verse" or "vamp", but that happens more often during the first and last choruses, while solo choruses usually keep the predictable structure that is necessary for jazz improvisation.

Problem statement. Assuming that the chord change usually happens simultaneously with the beat onset, we define a beat-level musical form. Let X be a certain song in the collection of lead sheets, then for each X we define a set of K sections that comprise its musical form: $\{S_i\}_{i=1}^{K}$. For each section S_i we define a set of d_i beats: $\{b_j^i\}_{j=1}^{d_i}$, where d_i is the duration of section S_i. For each beat we define a chord as a binary 12-dimensional chroma representation: $f\left(b_j^i\right) \rightarrow \{0,1\}^{12}$. Also we will use a distance function $D(b_{j1}^{i1}, b_{j2}^{i2})$ which represents how close in time these beats are (for example, consecutive beats b_j^i and b_{j+1}^i should have a small distance value).

In the recorded performance $\hat{X}$ there is a true sequence of beat onsets: $o_1, o_2, \ldots, o_N$, and $t(o_k)$ denotes the time between the start of the recording and the k-th beat onset. We assume that for each beat onset o_k there is a corresponding beat element of the musical form: $g\left(o_k\right) \rightarrow \{b_j^i\}$.

The first problem that we want to solve is *lead sheet alignment*: based on the audio signal a sequence of beats is estimated $\hat{o}_1, \hat{o}_2, \ldots, \hat{o}_{\hat{N}}$, and for each $\hat{o}_k$ we assign a beat element from the musical form: $\hat{g}\left(\hat{o}_k\right) \rightarrow \{b_j^i\}$. We seek $\{\hat{o}_k\}$ and $\hat{g}$ that have high beat recall and at the same time minimize the alignment error:

$$E\left(\hat{\boldsymbol{o}}, \hat{g} | X, \boldsymbol{o}, g\right) = \sum_{t(o_k)-t(\hat{o}_m)<\epsilon} D(g(o_k), \hat{g}(\hat{o}_m)). \quad (1)$$

Structural segmentation is a simple variant of this problem. In this case, we want to minimize the same alignment error, but calculated only for those beats where section boundaries are predicted (i.e. $o_k, \hat{o}_m$, where $g(o_k) = b_1^p$ and $\hat{g}(\hat{o}_m) = b_1^q$ for some p, q).

3. PROPOSED APPROACH

We propose an alignment approach that uses deep features extracted from the audio signal as observations in a Hidden Markov Model (HMM) designed after a corresponding lead sheet.

The first step is to extract features that contain some information regarding local harmonic context. Following [14, 6, 9, 15] we propose a multi-task convolutional-recurrent neural network (CRNN) to extract the beat onset and deep chroma information from the Harmonic Constant-Q Transform (HCQT, [16]) representations of the audio signal. The proposed CRNN is a modified and simplified version of network proposed in [9]. The details of the proposed architecture can be found in Figure 1, and the training details are discussed in Section 4. To create a comparable score representation we transform a symbolic lead sheet (a sequence of chords) into a score chromagram by concatenating 12-dimensional binary vectors, representing the pitch classes of chord tones with 1 and non-chord tones with 0.

Second, we design an HMM that emulates the performance generation process. We consider beat elements of musical form $\{b_j^i\}$ as hidden states in the HMM and the sequence of detected beat onsets $\hat{o}_i$ as observations.

Transition matrix. Within each section S_i we determine a natural "left-to-right" order of hidden states $\{b_j^i\}$ (for each state we have exactly one next state). Between sections we determine stochastic transitions, deriving $P^S = (p_{ik}^S)$ - probabilities of transition from section S_i to section S_k based on the information about musical form of the analyzed composition. For example, in "AABA" form the $A \rightarrow A$ transition probability is $\frac{2}{3}$, and the $A \rightarrow B$ transition probability is $\frac{1}{3}$. To define temporal evolution of the performance we introduce three parameters: p_{stay}, p_{step}, and p_{skip} that correspond to probabilities of staying in the same state, moving to the next state, and skipping the next state respectively. For states that are not at the end of a section these parameters comprise transition probability distributions. For the final and the penultimate beats of each section, parameters p_{step} and p_{skip} are distributed among following states proportionally to inter-section transition probabilities. Thus, the form of the achieved transition matrix is close to tridiagonal with additional non-zero values that correspond to transitions between sections.

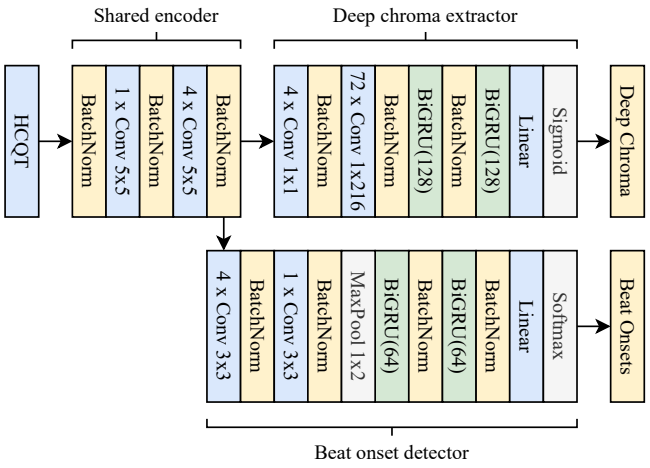

Figure 1: A multi-task CRNN for deep feature extraction.

Emission probabilities. We assign a categorical distribution for each possible chord. First we define a distance measure between 12-dimensional binary pitch class vectors, after that we calculate the distance between the chroma of a chosen chord and all 4096 possible 12-dimensional binary vectors. Then we apply a sigmoid function to these distances and normalize them. In our work we use the Hamming distance as a distance function, but it could be potentially replaced with a metric that more accurately captures harmonic affinity.

Finally, we set a uniform prior probability, allowing the analyzed segment to start at any possible location in a chorus. To decode the sequence of hidden states we use the Viterbi decoding algorithm. Since the key of the recording and the key of the lead sheet are often different, we transpose deep chroma features to 12 keys and perform decoding for each key independently, selecting the final alignment result by the highest decoding likelihood value.

4. EXPERIMENTAL SETUP

The **Jazz Audio-Aligned Harmony** (JAAH, [17]) dataset contains 113 audio files that were selected from two Smithsonian collections of jazz recordings and annotated specifically for training and evaluating automatic chord estimation models in the jazz music domain. State-of-the-art chord recognition models trained on polystylistic datasets show relatively weak performance on jazz recordings, inferior to the models trained on in-domain jazz recordings [17, 9],

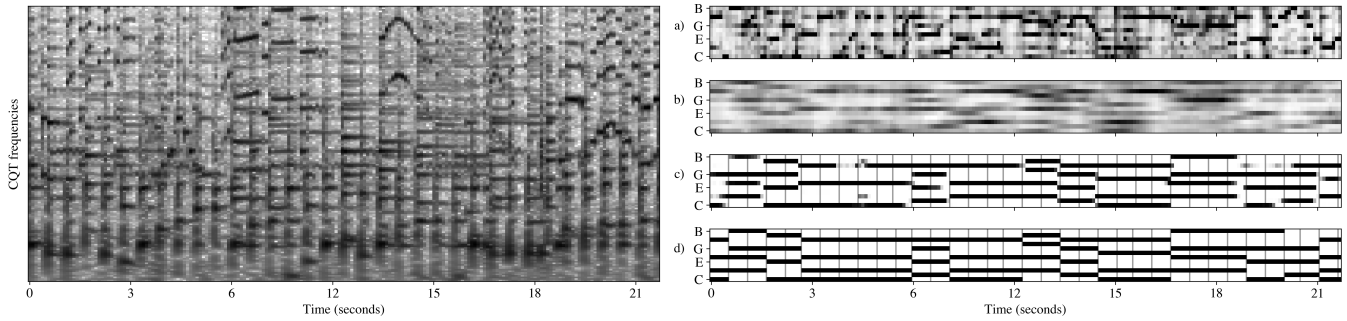

Figure 2: Comparison of features extracted from a 20-second excerpt of "Mean to Me" (JAAH dataset). Left: CQT features. Right: a) Raw chroma STFT features; b) Averaged chroma features (window size of two beats); c) Deep chroma features extracted with the proposed approach (vertical lines represent predicted beats); d) Ground-truth chromagram annotation from the JAAH dataset (vertical lines represent beat annotations).

which makes the JAAH dataset particularly important for jazz analysis. The annotations include semi-automatically detected beat positions, harmonic segmentation with chord symbols transcribed by qualified annotators, and a detailed structural segmentation. The structural segmentation is provided in a non-standardised form and requires manual editing to be included in an automatic training or evaluation pipeline. Following [9], we use this dataset to train the deep feature extractor proposed in Section 3.

The **Weimar Jazz Database** (WJD) [18] is comprised of 456 solo sections (audio fragments that contain an improvised solo) extracted from 340 jazz recordings. Annotations for these solo sections contain melodic transcriptions of the main instrument that includes note durations, beat positions, structural segmentation, and detailed recording metadata. This dataset also includes chord annotations, although the chords were not transcribed, but copied from lead sheets. Authors of this dataset mainly focused on creating a large collection of precise transcriptions of improvised solos, however the chorus segmentation and beat annotation make this dataset suitable for lead sheet alignment evaluation. Recently, the complete versions of the tracks from this dataset have been supplemented with additional structural annotation that became known as the **Jazz Structure Dataset** (JSD) [13]. Although the beat position information is not available for complete versions of recordings, the chorus-level structural segmentation can be used to measure the accuracy of estimated chorus boundaries. Note that the alignment test created from solo parts (WJD) captures the *beat-level* behaviour of the proposed system, while the chorus detection test on complete tracks (JSD) ignores any local errors in lead sheet alignment and can only be used to evaluate *chorus-level* structural segmentation.

The **IRealPro** dataset [19] is an open crowd-sourced collection of harmonic schemas for many popular jazz standards. A publicly available data release [20] contains 1186 entries, each of which consists of metadata (such as title, author, year, style, meter), structural information (names and the order of sections) and harmonic information (chord changes for each section). The content of this dataset is widely used by professional and amateur musicians via the iRealPro app.

In our implementation[1] we use HCQT features extracted from the 44.1 kHz sample rate audio files, using non-overlapping windows of 4096 samples ($\approx$ 93 ms), using the first two harmonics and a range of six octaves from C1 to C7. During training we apply chromatic augmentations, transposing the input to all 12 keys. We use segments of 100 consecutive frames and batches of 64 segments (input dimensions: $64 \times 2 \times 100 \times 216$). 110 tracks from the JAAH dataset are used as a training dataset and 3 tracks as a validation set ("Mean to Me", "Lady Bird" and "Blue Seven"). We use a sum of cross-entropy loss functions for deep chroma extraction and beat onset detection with the Adam optimizer (learning rate: 0.001, weight decay: 0.01). Early stopping is used, where training is stopped when the validation loss does not improve over 10 epochs. For Viterbi decoding we average deep chroma representations over the duration of one beat (as detected by the proposed system) and binarize the result using threshold 0.5. To compile the transition matrix we use parameters $p_{stay} = 0.1$, $p_{step} = 0.8$, and $p_{skip} = 0.1$. These probabilities were chosen empirically to make the solution more robust to errors in beat tracking and they are loosely based on the performance evaluation of the beat onset detector. In our work we are using multiple datasets with harmonic annotations, so it is important to have comparable representations of harmonic context. We convert the chords from the IRealPro dataset to the widely used notation proposed by Harte et al. [21] using parsers and grammars developed for the ChoCo corpus [22].

We use solo sections from the WJD [18] to measure the beat-level alignment quality. Since the WJD includes both ground-truth beat information and chorus segmentation, we use this annotation to match each beat with an audio timestamp and with a beat number in the current chorus. We calculate beat-level accuracy by comparing this annotation with the result of Viterbi decoding using three levels of tolerance (2 beats, 4 beats and 8 beats), motivated by the fact that improvised melodic phrases are not always perfectly aligned with harmonic context. This test has an important limitation that it does not include any modulations or other significant deviations from musical form since it consists mostly of solo fragments of commercially available jazz recordings.

To measure structural segmentation performance we utilise the chorus segmentation that is available for complete tracks of the WJD provided in the JSD. Annotations for complete tracks do not contain beat onset information so we evaluate chorus detection generated by the Viterbi decoding. We consider that the chorus is detected correctly if its predicted boundaries are within the tolerance interval of ground truth timestamps. This logic is used to calculate precision, recall, and F-measure for chorus boundary detection.

[1] https://github.com/shanin/waspaa2023-lead-sheet-alignment

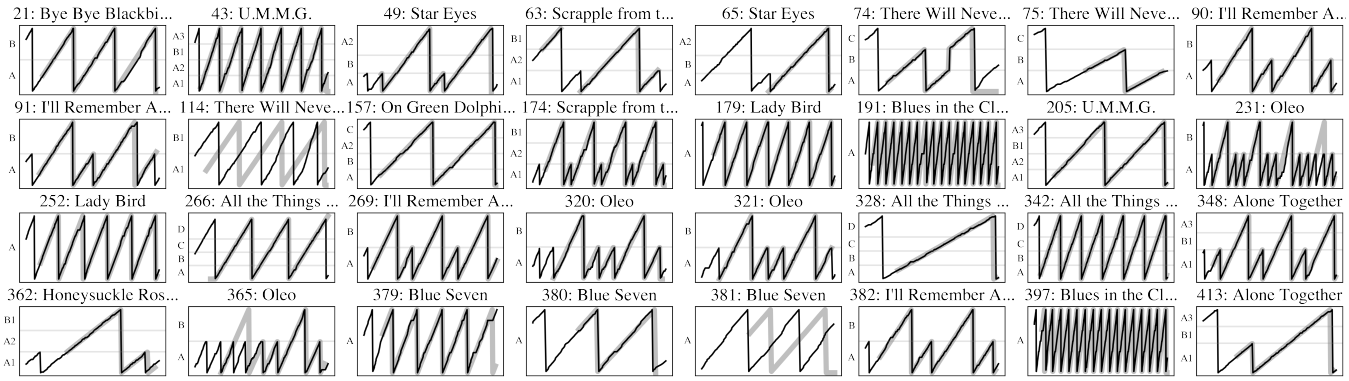

Figure 3: Lead sheet alignment results on a subset of WJD. Section names in IRealPro were manually matched with section names in WJD for this experiment (in favour of the WJD naming convention). Each plot has time on the x-axis and HMM state on the y-axis. The gray background line represents the beat ground truth annotation from WJD.

	τ	C	M	CA	**MA**	CAG
JSD (f1)	1s	0.19	0.26	0.21	0.30	–
	2s	0.24	0.33	0.26	0.39	–
	3s	0.27	0.36	0.28	0.43	–
WJD (acc.)	2b	0.26	0.35	0.26	0.46	0.69
	4b	0.47	0.59	0.48	0.66	0.75
	8b	0.50	0.63	0.50	0.71	0.77

Table 1: Performance (F-measure with tolerance τ) of the proposed multi-task CRNN ("M") always exceeds that of a baseline deep chroma extractor ("C") trained only on the chroma prediction task and used in combination with the state-of-the-art beat tracker [23]. "A" models were trained using chromatic augmentations. For the WJD test we also use ground-truth beat annotation ("G") to evaluate sensitivity to beat tracking errors.

5. RESULTS

We found that the beat tracking performance significantly impacts decoding quality, while chromatic augmentations improve the quality of deep chroma representations (Table 1). We provide a brief error analysis for a part of this test set that consists of 32 solo sections (Figure 3). Most errors result from similarity of chord sequences between sections. The two sections of solo 114 are 60% identical, leading to confusion between them. Solos 231 and 365 follow the commonly known "rhythm changes" harmonic structure, where the last two bars of sections labelled A and B share the same harmonic function, which is also identical to the first half of A; the proposed system mislabels one of the B sections in each of these samples, while correctly labeling other sections. Solo 381 was decoded entirely wrongly due to the fact that the highest decoding likelihood was in the wrong key. All other samples in this experiment were decoded without structural errors.

6. DISCUSSION AND CONCLUSION

There are several significant differences between our approach and [12]. First, we use deep chroma audio features instead of STFT chromagrams. We find deep chroma representations to be closer to the desired annotation than beat-averaged chromagrams that were previously used (see Figure 2). Second, in [12] the complete lead sheets were used to create score chromagrams, which included not only chords, but also the melody. Our approach uses only chords to create score chromagrams, which is also important since melodies are protected by copyright, but chord changes are copyright-free, thus easier to obtain. Finally, [12] is not scalable: only three compositions were manually prepared as a test dataset, and our approach using automatic score chroma generation was tested on 140 compositions and could be scaled to work with large corpora of jazz recordings. In order to compare the performance of the proposed system with [12] we analyzed four commercial recordings of "Without a Song" (Figure 4), that were mostly mislabeled by their approach. Our system shows clearly better performance predicting correct structure despite harmonic variations, metric alterations ("half-time feel") and deviations from form (extra "vamp" sections).

Regarding the JSD and WJD tests that were used in this work, we want to address the issue of consistency of section boundary annotation in publicly available datasets. Performance analysis of the proposed system showed that in many cases the annotated section boundaries were different from conventional lead sheets: for example, quite often annotated sections were in fact "AA" and "BA" parts of the "AABA" form. This directly affects the evaluation results provided in Table 1 and explains the difference between the whole WJD test and the results presented in Figure 3 where section and chorus segmentation were rectified manually.

Summarizing the above, in this paper we proposed a new lead sheet alignment algorithm based on deep chroma features and Viterbi decoding. It outperforms previously known approaches and could be used for the automatic annotation of large-scale jazz recording corpora.

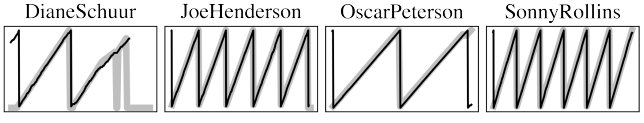

Figure 4: Performance of the proposed system on four recordings of "Without a Song" that were previously analysed in [12].

7. REFERENCES

[1] M. McVicar, R. Santos-Rodriguez, Y. Ni, and T. de Bie, "Automatic chord estimation from audio: A review of the state of the art," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 556–575, 2014.

[2] J. Pauwels, K. O'Hanlon, E. Gómez, and M. B. Sandler, "20 years of automatic chord recognition from audio," in *20th International Society for Music Information Retrieval Conference*, 2019, pp. 54–63.

[3] D. Odekerken, H. V. Koops, and A. Volk, "Improving audio chord estimation by alignment and integration of crowdsourced symbolic music," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, pp. 141–155, 2021.

[4] 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, 2010.

[5] 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, 2011.

[6] B. McFee and J. P. Bello, "Structured training for large-vocabulary chord recognition," in *International Society for Music Information Retrieval Conference*, 2017, pp. 188–194.

[7] T.-P. Chen and L. Su, "Harmony transformer: Incorporating chord segmentation into harmony recognition," in *20th International Society for Music Information Retrieval Conference*, 2019, pp. 259–267.

[8] V. Eremenko, E. Demirel, B. Bozkurt, and X. Serra, "Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research." in *19th International Society for Music Information Retrieval Conference*, 2018, pp. 483–490.

[9] G. Durán and P. de la Cuadra, "Transcribing lead sheet-like chord progressions of jazz recordings," *Computer Music Journal*, vol. 44, no. 4, pp. 26–42, 2020. [Online]. Available: https://doi.org/10.1162/comj_a_00579

[10] Y. Wu, T. Carsault, and K. Yoshii, "Automatic chord estimation based on a frame-wise convolutional recurrent neural network with non-aligned annotations," in *27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[11] M. Grachten, M. Gasser, A. Arzt, and G. Widmer, "Automatic alignment of music performances with structural differences," in *14th International Society for Music Information Retrieval Conference*, A. de Souza Britto Jr., F. Gouyon, and S. Dixon, Eds., 2013, pp. 607–612. [Online]. Available: http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/158_Paper.pdf

[12] Z. Duan and B. Pardo, "Aligning semi-improvised music audio with its lead sheet." in *International Society for Music Information Retrieval Conference*, 2011, pp. 513–518.

[13] S. Balke, J. Reck, C. Weiß, J. Abeßer, and M. Müller, "Jsd: A dataset for structure analysis in jazz music," *Transactions of the International Society for Music Information Retrieval*, vol. 5, no. 1, 2022.

[14] F. Zalkow and M. Müller, "Using weakly aligned score-audio pairs to train deep chroma models for cross-modal music retrieval," in *21st International Society for Music Information Retrieval Conference*, J. Cumming, J. H. Lee, B. McFee, M. Schedl, J. Devaney, C. McKay, E. Zangerle, and T. de Reuse, Eds., 2020, pp. 184–191. [Online]. Available: http://archives.ismir.net/ismir2020/paper/000023.pdf

[15] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," in *17th International Society for Music Information Retrieval Conference*, M. I. Mandel, J. Devaney, D. Turnbull, and G. Tzanetakis, Eds., 2016, pp. 37–43. [Online]. Available: https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/178_Paper.pdf

[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 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 63–70. [Online]. Available: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/85_Paper.pdf

[17] V. Eremenko, E. Demirel, B. Bozkurt, and X. Serra, "Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research," in *19th International Society for Music Information Retrieval Conference*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 483–490. [Online]. Available: http://ismir2018.ircam.fr/doc/pdfs/206_Paper.pdf

[18] M. Pfleiderer, K. Frieler, J. Abeßer, W.-G. Zaddach, and B. Burkhart, *Inside the Jazzomat: New Perspectives for Jazz Research*. Schott Campus, 2017.

[19] D. Shanahan, Y. Broze, and R. Rodgers, "A diachronic analysis of harmonic schemata in jazz," in *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music*, 2012, pp. 909–917.

[20] D. Shanahan and Y. Broze, "iRealPro Corpus of Jazz Standards (v1.0) [Data set]," 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3546040

[21] C. Harte, M. B. Sandler, S. A. Abdallah, and E. Gómez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *6th International Conference on Music Information Retrieval*, 2005, pp. 66–71. [Online]. Available: http://ismir2005.ismir.net/proceedings/1080.pdf

[22] J. de Berardinis, A. Meroño-Peñuela, A. Poltronieri, and V. Presutti, "ChoCo: A chord corpus and a data transformation workflow for musical harmony knowledge graphs," in *Manuscript under review*, 2023.

[23] S. Böck, M. E. Davies, and P. Knees, "Multi-task learning of tempo and beat: Learning one to improve the other." in *20th International Society for Music Information Retrieval Conference*, 2019, pp. 486–493.