Automatic Chord Estimation Based on a Frame-wise Convolutional Recurrent Neural Network with Non-Aligned Annotations

Yiming Wu* Tristan Carsault† Kazuyoshi Yoshii*
*Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
Email: {wu, yoshii}@sap.ist.i.kyoto-u.ac.jp
†IRCAM, CNRS, Sorbonne Université, UMR 9912 STMS, Paris, France
Email: carsault@ircam.fr

***Abstract*—This paper describes a weakly-supervised approach to Automatic Chord Estimation (ACE) task that aims to estimate a sequence of chords from a given music audio signal at the frame level, under a realistic condition that only non-aligned chord annotations are available. In conventional studies assuming the availability of time-aligned chord annotations, Deep Neural Networks (DNNs) that learn frame-wise mappings from acoustic features to chords have attained excellent performance. The major drawback of such frame-wise models is that they cannot be trained without the time alignment information. Inspired by a common approach in automatic speech recognition based on non-aligned speech transcriptions, we propose a two-step method that trains a Hidden Markov Model (HMM) for the forced alignment between chord annotations and music signals, and then trains a powerful frame-wise DNN model for ACE. Experimental results show that although the frame-level accuracy of the forced alignment was just under 90%, the performance of the proposed method was degraded only slightly from that of the DNN model trained by using the ground-truth alignment data. Furthermore, using a sufficient amount of easily collected non-aligned data, the proposed method is able to reach or even outperform the conventional methods based on ground-truth time-aligned annotations.**

***Index Terms*—Automatic chord estimation, forced alignment, HMM, CNN, and RNN.**

I. Introduction

Harmonic progression is an important property of music in western music theory, which provides rich information about the characteristics of the musical work. In the Music Information Retrieval (MIR) community, Automatic Chord Estimation (ACE) [1] has been a long-lasting research theme. Modelling and automating the estimation process has still been a challenging task due to the complexity of music signals and the implicit relationships between the chord labels.

A standard approach to ACE is to extract acoustic features that represent the harmonic characteristics of music signals and then train some machine learning models in a data-driven manner. Although the performance of ACE has recently been improved remarkably [2], conventional studies stand on the assumption that time-aligned chord annotations are available as supervised data. Producing such annotations is costly, since manual chord annotation requires careful judgments based on knowledge on musical theories, and time alignments have to be carefully given to each chord symbol. This makes it difficult to extend the scale of supervised data efficiently, and the limitation becomes especially distinct when Deep Neural Networks (DNNs) are introduced for estimating the posteriors.

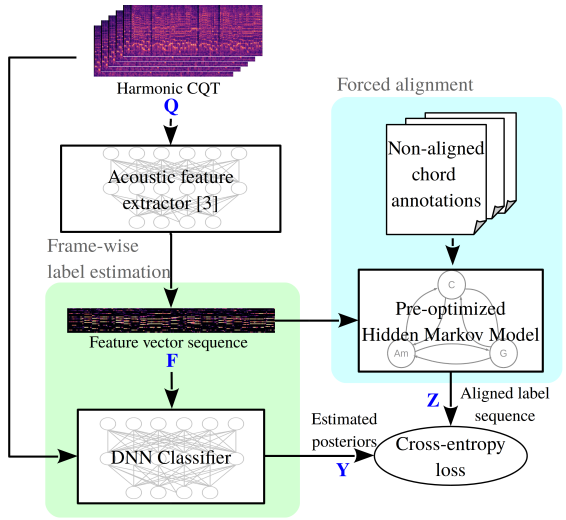

Fig. 1. The proposed training strategy based on an HMM for forced alignment and a DNN for automatic chord estimation.

In automatic speech recognition (ASR), which has a similar task formulation, a recognition model is typically trained on *weaker* labelled data. Only a sequence of symbols (phonemes or words) is given for the audio signal of each utterance without any time alignment information. In order to train a DNN model such as the HMM-DNN hybrid model [3] into the framework of ASR, the annotated symbols are firstly aligned to frame-wise acoustic feature sequences with a trained HMM (forced-alignment). Then a DNN is trained for accurately estimating the label posteriors at each frame.

Inspired by such an approach to ASR, in this paper we propose an ACE method based on a two-step training strategy (Fig. 1). Our method extracts frame-wise chroma vectors from music signals by using a state-of-the-art DNN-based feature extractor [4]. In the training phase, an HMM that represents chord labels as latent and feature vectors as observed variables is trained for estimating the time alignment between a feature

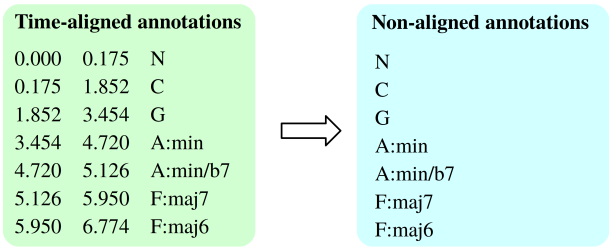

Fig. 2. *Time-aligned* and *non-aligned* chord labels. In *time-aligned* labels, the start and end timing of each chord are annotated in seconds.

sequence and the corresponding non-aligned chord annotations (Fig. 2). A frame-wise classifier based on a Convolutional Recurrent Neural Network (CRNN) is then trained on the time-aligned pairs of feature sequences and chord label sequences. In the test phase, the trained classifier is used for estimating the posterior probabilities of chord labels at each frame from a given music signal and another HMM is used for estimating the optimal path of chord labels.

The main contribution of this paper is to show that non-aligned chord annotations are useful for effectively training machine-learning based ACE models. Furthermore, we show that by using a sufficient amount of training data with non-aligned annotations, the trained DNN model can reach or even outperform the DNN models trained with only ground-truth time-aligned chord annotations. To the best of our knowledge, this is the first attempt in ACE to train DNN model on non-aligned chord annotations.

II. Related Work

Some common methodologies have been developed in order to extract effective harmonic features from the audio signal. The *chroma vector*, which indicates the relative intensities of chromatic pitch classes in each audio frame [5], is the most representative and has been widely employed. Based on the characteristics of harmonic structures in the frequency domain, many effective techniques to extract chroma features have been proposed (*e.g.*, [6], [7]). Furthermore, data-driven approaches have recently been considered to be promising [4], [8], [9].

The core of ACE is to model the translation process from a frame-level domain (harmonic features) to a symbol-level domain (chord labels). Conventionally, the relationships between a feature sequence and a chord label sequence is modeled by a Hidden Markov Model (HMM) [6], [10]. However, since Humphrey and Bello [11] proposed a Convolutional Neural Network (CNN) model, discriminative methods that use DNNs for directly estimating a sequence of the posterior probabilities of chord labels have gained a lot of attention [12], [13],

A common approach is to estimate the posterior probabilities at the frame level directly from low-level time-frequency representations such as Short-Time Fourier Transform (STFT) or Constant-Q Transform (CQT) spectrograms rather than hand-crafted features. A musically-meaningful design of the output of a DNN based on chord theory [14], [15] and an even-chance training scheme [16] can improve the performance of ACE, especially when a large chord vocabulary is used and the frequencies of chords are heavily biased. Another approach is to estimate the posterior probabilities at the symbol level. Deng and Kwok [17] proposed a method that uses a pre-trained HMM for estimating chord boundaries and then uses a DNN for estimating a chord for each segment consisting of multiple frames. This method is better at estimating seventh chords than frame-level ACE models.

The conventional methods mentioned above make use of carefully time-aligned chord annotations as supervising data. A number of time-aligned annotation datasets have been released for training and evaluating those models. In the datasets used in MIREX evaluation, for example, more than 1000 annotated pieces are available in total. Considering the other successful domains such as ASR, the amount of the training data is still comparatively small. It is thus natural to expect that there is still room for improving the performance of ACE if a larger amount of training data would be available.

III. Proposed Method

This section describes the proposed method of ACE based on a CRNN trained from non-aligned chord annotations.

A. Problem Specification

Let $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ be a frame-level sequence of acoustic features extracted from a music signal, where $\mathbf{x}_n \in \mathbb{R}^D$ is a D-dimensional feature vector and N is the number of time frames. Let $\mathbf{Z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_N\}$ be a sequence of chord labels, where $\mathbf{z}_n \in \{0,1\}^K$ is a one-hot vector indicating a chord label in a chord vocabulary of size K. The task of statistical ACE is defined as follows:

$$\mathbf{Z}^* = \underset{\mathbf{Z}}{\operatorname{argmax}}\, p(\mathbf{Z}|\mathbf{X}), \tag{1}$$

where $\mathbf{Z}^*$ is the optimal sequence of chord labels that maximizes the posterior probability $p(\mathbf{Z}|\mathbf{X})$.

In this task, we need to estimate both chord symbols and boundaries. Typically, $p(\mathbf{Z}|\mathbf{X})$ is evaluated by aligning $\mathbf{Z}$ with $\mathbf{X}$ at the frame level and estimating the frame-wise posterior probability $p(\mathbf{z}_n|\mathbf{x}_n)$. This is the main reason why time alignment information is required for training an ACE model.

As a primordial step, a DNN-based feature extractor proposed by Wu [4] is used for extracting a sequence of chroma vectors $\mathbf{F} = \{\mathbf{f}_1, \cdots, \mathbf{f}_N\}$ from the 5-channel Harmonic CQT (HCQT) [18] representation $\mathbf{Q} = \{\mathbf{q}_1, \cdots, \mathbf{q}_N\}$ of a music signal, where $\mathbf{f}_n \in \mathbb{R}^{36}$ is a 36-dimensional vector representing the relative intensities of the 12 chromatic pitch classes in lower, middle, and higher frequency ranges, respectively, and $\mathbf{q}_n \in \mathbb{R}^{5\times 96}$ is the HCQT coefficients. Each $\mathbf{x}_n$ contains the two representations of the frame n, *i.e.*, $\mathbf{x}_n = \{\mathbf{f}_n, \mathbf{q}_n\}$.

B. HMM-Based Forced Alignment

To synchronize non-aligned chord annotations with a feature sequence $\mathbf{F}$ at the frame level, we formulate a frame-wise HMM that represents $\mathbf{Z}$ and $\mathbf{F}$ as latent and observed variables, respectively, as follows:

$$p(\mathbf{F}, \mathbf{Z}) = \prod_{n=1}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}) p(\mathbf{f}_n|\mathbf{z}_n), \tag{2}$$

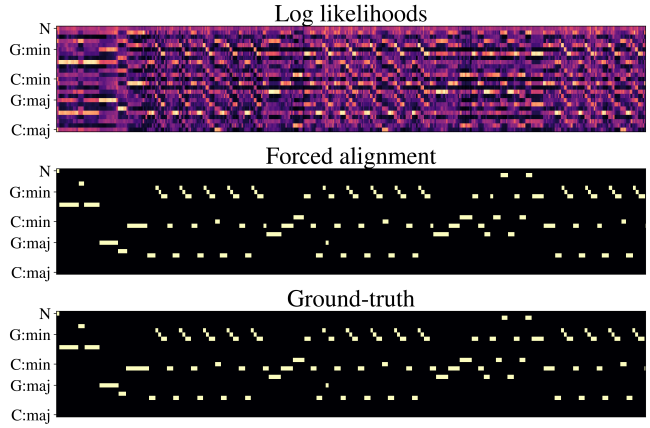

Fig. 3. An example of the HMM-based forced alignment. From top to bottom, a sequence of the likelihoods over chord labels, the optimal path estimated by the HMM, the ground-truth time-aligned sequence of chord labels are shown. Chord labels other than major and minor triads are omitted.

where $\mathbf{z}_0$ is a dummy state. In this paper we use a von Mises-Fisher distribution as the emission distribution as follows:

$$p(\mathbf{f}_n|\mathbf{z}_n) = \prod_{k=1}^{K} \mathrm{vMF}(\bar{\mathbf{f}}_n|\boldsymbol{\mu}_k, \lambda_k)^{z_{nk}}, \tag{3}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^{36}$ satisfying $||\boldsymbol{\mu}_k|| = 1$ and $\lambda_k > 0$ are a mean direction vector and a concentration parameter, respectively. $\bar{\mathbf{f}}_n$ is obtained by scaling $\mathbf{f}_n$ to satisfy $||\bar{\mathbf{f}}_n|| = 1$. The transition distribution is given by

$$p(\mathbf{z}_n|\mathbf{z}_{n-1}) = \prod_{k=1}^{K}\prod_{k'=1}^{K} \pi_{kk'}^{z_{n-1,k}z_{nk'}}, \tag{4}$$

where $\pi_{kk'}$ is the transition probability from chord k to chord k'. Since this HMM is a frame-level model, the self-transition probabilities are considered to be close to 1.

Given a feature sequence $\mathbf{F}$ with non-aligned chord annotations (Fig. 2), we estimate a chord sequence $\mathbf{Z}$ that maximizes the posterior probability $p(\mathbf{Z}|\mathbf{F}) \propto p(\mathbf{F}, \mathbf{Z})$. Since the chord transitions in $\mathbf{Z}$ are specified by the annotations, in this paper we train a left-to-right HMM by using a Viterbi algorithm. After the HMM is initialized, the HMM searches the optimal state transitions for $\mathbf{F}$ and updates the parameters $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\pi}$ in a way of maximum likelihood estimation. These two steps are iterated for several times until convergence. An example of such forced alignment is shown in Fig. 3. Since the likelihood $p(\mathbf{f}_n|\mathbf{z}_n)$ is ambiguous, it is difficult to estimate a sequence of chord labels with the ordinary unconstrained Viterbi algorithm. Nonetheless, when the order of chord symbols is constrained, the HMM can perform the forced alignment accurately. This fact leads to our motivation to train a powerful DNN for ACE based on the annotations aligned by the HMM instead of directly using the HMM for ACE.

C. CRNN-Based Chord Estimation

We use a Convolutional Recurrent Neural Network (CRNN) for estimating a sequence of the posterior probabilities of chord labels $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$ from $\mathbf{X}$. In this study, the chord vocabulary consists of six triad types such as *maj*, *min*, *diminished*, *augmented*, *sus2*, and *sus4*. The output size of the last layer is thus 73 (72 triads plus a *no-chord* label), *i.e.*, $\mathbf{y}_n \in \mathbb{R}^{73}$. As shown in Table I, the two representations $\mathbf{F}$ and $\mathbf{Q}$ are processed separately at the bottom layers, with a CRNN for $\mathbf{Q}$ and a multi-layer bi-directional LSTM network for $\mathbf{F}$. The convolutional layers dealing with the spectrogram $\mathbf{Q}$ consists of 10 stacked residual blocks [19] with two convolutional layers. The outputs of the two networks are concatenated at the top layer to calculate the final output $\mathbf{Y}$.

TABLE I
CONFIGURATION OF THE PROPOSED CRNN.

Input: $\mathbf{F}(N \times 36)$	**Input:** $\mathbf{Q}(N \times 5 \times 96)$
Bi-directional LSTM 128 units $\times 3$ layers	Convolution $64 \times 15 \times 3$
	Convolution $256 \times 3 \times 3$
	max pooling 3×3
	Residual block: ([Convolution $256 \times 5 \times 5$] $\times 2$) $\times 10$
	Max pooling 3×4
	Convolution $256 \times 15 \times 7$
	Bi-directional LSTM 128 units $\times 4$ layers
Concatenate	
Fully-connected 256×73	

*The output of each convolutional layer is activated with the ReLU function and then batch-normalized.

The CRNN is trained in a standard way. More specifically, the cross-entropy loss between the target binary sequence $\mathbf{Z}$ and the output sequence $\mathbf{Y}$ is iteratively minimized by using Adam optimizer [20]. To improve the generalization capability of the CRNN, we propose a data augmentation technique. In each iteration, the feature sequence $\mathbf{F}$ and the corresponding target label sequence $\mathbf{Z}$ are pitch-shifted by a random number (up to 12) of semitones. Note that the input spectrogram $\mathbf{Q}$ is not shifted. This operation is expected to act as a strong regularization during the training process, encouraging the convolutional layers to focus on the *shapes* of chords (*i.e.*, chord types) rather than actual pitch information when processing the spectrogram.

D. HMM-Based Postprocessing

The final step of the proposed ACE method is to estimate a sequence of chord symbols with boundary information from a sequence of the frame-wise posterior probabilities of chord labels $\mathbf{Y}$ estimated by the CRNN. This is a common smoothing step of existing HMM-based systems [10]. More specifically, we formulate an HMM that has chord labels $\mathbf{Z}$ as latent variables and outputs the posterior probabilities $\mathbf{Y}$. The transition probabilities are given by $\boldsymbol{\pi}$. Using the Viterbi algorithm, the optimal sequence $\mathbf{Z}^*$ can be obtained.

An example of the estimated posterior probabilities $\mathbf{Y}$ and the final outputs $\mathbf{Z}^*$ is shown in Fig 4. Short-duration outliers cannot be avoided if the label with the maximum probability is selected in each frame. The postprocessing step can reduce these outliers and obtain more natural results.

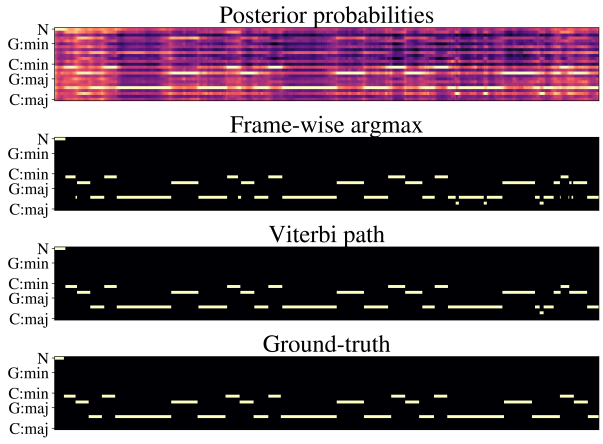

Fig. 4. An example of the HMM-based postprocessing. From top to bottom, a sequence of the posterior probabilities over chords estimated by the CRNN, a sequence of chord labels obtained by the frame-wise argmax operation, that obtained by the HMM, and the ground-truth label sequence.

IV. Evaluation

This section reports a comparative experiment conducted for evaluating the proposed weakly-supervised method in comparison with the completely supervised method.

A. Experimental Conditions

The frame-level accuracy of ACE for each musical piece is measured by comparing the ground-truth and estimated chord sequences with *mir_eval* library [21]. The average performance is calculated over the piece-wise accuracies weighed by the numbers of time frames of the musical pieces. We define two different chord vocabularies: the *Majmin* vocabulary that consists of major (*maj*) and minor (*min*) triads plus no chord (25 classes), and the *Triads* vocabulary that additionally includes augmented (*aug*), suspended (*sus2* and *sus4*), diminished (*dim*) triads plus no chord (73 classes).

For evaluation, we use both the Isophonics dataset consisting of 220 songs of Queen, Zweieck, and The Beatles [22] and the RWC popular music dataset consisting of 100 Japanese popular songs [23] with chord annotations, *i.e.*, 320 songs in total. All the annotations are time-aligned so that the performance can be measured. To investigate the impact of increasing the amount of supervising data for the proposed method, we additionally use as a larger dataset with non-aligned annotations a subset of the Mcgill Billboard dataset [24] consisting of 731 pieces that are disjoint with the Isophonics and RWC datasets. Using these datasets, we design three different training configurations. In any configuration, the performance is measured on the Isophonics+RWC dataset by performing 5-fold cross validation.

1) Isophonics+RWC (time-aligned): The CRNN is trained on the ground-truth *time-aligned* annotations without using the HMM-based forced alignment.
2) Isophonics+RWC (non-aligned): This is the same as the Isophonics+RWC configuration except that the CRNN is trained on the *non-aligned* chord annotations by using the HMM-based forced alignment.
3) Isophonics+RWC+Billboard (non-aligned): The CRNN is trained on a *larger* amount of *non-aligned* chord annotations (Isophonics+RWC+Billboard dataset) by using the HMM-based forced alignment. In each fold, 80% of the Isophonics+RWC dataset and the Billboard dataset are used for training the CRNN.

For comparison, we test two existing ACE methods:

a) Chordino [6]: This is a method based on an HMM using the NMF-chroma feature, which is available as a VAMP plugin. As a baseline, we use this method with a pre-trained model to analyze the Isophonics+RWC datasets without performing cross-fold validation.
b) CNN-CRF [13]: This is the state-of-the-art method based on a fully convolutional acoustic model combined with a conditional random field (CRF), We use a reference code implemented by the author according to the original paper. Since the method estimates *major/minor* triads only, the scores for *Triads* are not available.

TABLE II
Experimental results on the Isophonics and RWC datasets (%).

	Isophonics		RWC	
	Majmin	*Triads*	*Majmin*	*Triads*
Isophonics+RWC (time-aligned annotations)	83.31	79.92	**82.24**	76.73
Isophonics+RWC (non-aligned annotations)	82.49	79.44	80.53	76.40
Isophonics+RWC+Billboard (non-aligned annotations)	83.46	**79.94**	82.21	**76.74**
Chordino [6]	72.21	74.24	78.78	73.66
CNN-CRF [13]	**84.22**	N/A	81.68	N/A

B. Experimental Results

The experimental results are listed in Table II. Comparing the first two rows, we observe that when the Isophonics+RWC dataset is used as training data, the weakly supervised method based on the forced-aligned annotations only slightly underperformed the supervised method based on the ground-truth annotations. When the Isophonics+RWC+Billboard dataset with non-aligned annotations are used as training data, the weakly supervised method performed even better than the completely supervised method. This shows that the performance of ACE obtained by the proposed method depends on the size of non-aligned training data, and that the performance limitation of the HMM-based forced alignment can be mitigated by collecting a sufficient amount of weakly-annotated data.

In all cases, the average accuracy of the forced alignment is around 88.5% for the Isophonics dataset and around 86.5% for the RWC dataset. Since the CRNN is trained to emulate the alignment behavior of the HMM, it is natural to think that the accuracy of ACE is upper bounded by the forced alignment accuracy. In fact, interestingly, the weakly-supervised method is still competitive to the state-of-the-art conventional models in estimating chord sequences for unseen data.

The performance differences between the two metrics (*Majmin* and *Triads*) are almost the same in all methods. For the

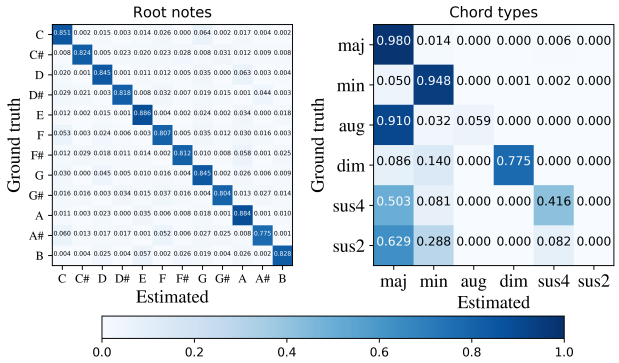

Fig. 5. The confusion matrix over the 12 root notes and that over the 6 chord types. In the left matrix, the comparison is done over all frames, while in the right matrix the comparison is done for time frames where the chord roots are correctly estimated.

Isophonics and RWC datasets, the performance differences are around 3% and around 5%, respectively. This indicates that the ability of classifying rarely-used triads other than *maj* and *min* is scarcely improved, even though the amount of training data is increased and the regularization technique is used for the training. The confusion matrices are shown in Fig. 5. The CRNN is excellent at estimating the frequently-used *maj* and *min* triads, while unreliable in estimating the other triads. In the whole dataset, the total duration of *aug* or *sus2* is nearly 10 minutes, *dim* is around 20 minutes, and *sus4* is up to 1.4 hours. We find that the CRNN trained in the proposed strategy is relatively good at estimating *dim*, is slightly worse for *sus4*, and fails to recognize *aug* and *sus2*.

The main limitation of the common approach to ACE based on the one-octave chromatic representation is that some chords are hard to distinguish in principle. The pitch classes of *sus2* chords, for example, are a subset of those of *9th* chords, which are regarded as *maj* chords in the standard *Triads* condition. The same is true for *sus4* and *11th* chords. A reason why the accuracy on *dim* is better than that on *sus4*, even though *dim* appear much less frequently, would be that *dim* and *aug* chords are less ambiguous. In addition to the heavily biased distribution of chord types, such ambiguity is considered to affect the accuracy when the traditional flat classification scheme is used, as suggested in [25]. One solution is to fully consider a wider range of octaves instead of compressing frequency spectra into one octave. In chord theory, rare triads are considered to play unique roles in chord progressions, *e.g.*, a *sus4* chord implies that a *maj* chord with the same root note comes next. A language model of chord progressions would be effective to solve the ambiguity.

V. Conclusion

This paper proposed a two-step training method that trains a CRNN-based chord estimation model with non-aligned chord annotations, based on an HMM-based forced alignment model that recovers the time alignments of the annotations. We experimentally showed that the model trained in the proposed method is competitive to the ones trained on time-aligned annotations, and can even outperform the completely supervised model by using a larger amount of non-aligned data. Our study opens up a door to make full use of non-aligned data that can be collected easily from the Web, but has not conventionally been used. To make the estimation results more accurate and musically natural, we plan to integrate a symbol-wise chord language model with the frame-wise estimation model.

Acknowledgment

This work was supported in part by JST ACCEL No. JPMJAC1602, JSPS KAKENHI No. 16H01744, No. 16J05486, and No. 19H04137, Kayamori Foundation, and the Kyoto University Foundation.

References

[1] Meinard Müller, *Fundamentals of music processing: Audio, analysis, algorithms, applications*, Springer, 2015.
[2] M. McVicar *et al.*, "Automatic chord estimation from audio: A review of the state of the art," *IEEE TASLP*, vol. 22, no. 2, pp. 556–575, 2014.
[3] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
[4] Y. Wu and W. Li, "Automatic audio chord recognition with MIDI-trained deep feature and BLSTM-CRF sequence decoding model," *IEEE TASLP*, vol. 27, no. 2, pp. 355–366, 2019.
[5] T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *ICMC*, 1999, pp. 464–467.
[6] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," in *ISMIR*, 2010, pp. 135–140.
[7] E. Gómez, *Tonal description of music audio signals*, Ph.D. thesis, Universitat Pompeu Fabra, 2006.
[8] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," in *ISMIR*, 2016, pp. 37–43.
[9] E. J. Humphrey *et al.*, "Learning a robust Tonnetz-space transform for automatic chord recognition," in *ICASSP*, 2012, pp. 453–456.
[10] R. Chen *et al.*, "Chord recognition using duration-explicit hidden Markov models," in *ISMIR*, 2012, pp. 445–450.
[11] E. J. Humphrey and J. P. Bello, "Rethinking automatic chord recognition with convolutional neural networks," in *ICMLA*, 2012, vol. 2, pp. 357–362.
[12] S. Sigtia *et al.*, "Audio chord recognition with a hybrid recurrent neural network," in *ISMIR*, 2015, pp. 127–133.
[13] F. Korzeniowski and G. Widmer, "A fully convolutional deep auditory model for musical chord recognition," in *MLSP*, 2016, pp. 13–16.
[14] B. Mcfee and J. P. Bello, "Structured training for large-vocabulary chord recognition," in *ISMIR*, 2017, pp. 188–194.
[15] T. Carsault *et al.*, "Using musical relationships between chord labels in automatic chord extraction tasks," in *ISMIR*, 2018, pp. 18–25.
[16] J. Deng and Y. K. Kwok, "Large vocabulary automatic chord estimation with an even chance training scheme," in *ISMIR*, 2017, pp. 531–536.
[17] J. Deng and Y. K. Kwok, "A hybrid Gaussian-HMM-deep learning approach for automatic chord estimation with very large vocabulary," in *ISMIR*, 2016, pp. 812–818.
[18] R. M. Bittner *et al.*, "Deep salience representations for F0 tracking in polyphonic music," in *ISMIR*, 2017, pp. 63–70.
[19] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
[21] C. Raffel *et al.*, "mir_eval: A transparent implementation of common MIR metrics," in *ISMIR*, 2014.
[22] C. Harte, *Towards automatic extraction of harmony information from music signals*, Ph.D. thesis, Queen Mary University of London, 2010.
[23] M. Goto *et al.*, "RWC music database: Popular, classical, and jazz music databases," in *ISMIR*, 2002, pp. 287–288.
[24] J. A. Burgoyne *et al.*, "An expert ground truth set for audio chord recognition and music analysis," in *ISMIR*, 2011, pp. 633–638.
[25] E. J. Humphrey and J. P. Bello, "Four timely insights on automatic chord estimation," in *ISMIR*, 10 2015, pp. 673–679.