Accuracy Improvement of Automatic Chord Recognition with Source Separation Preprocessing

Ayumu Mitoma* and Ken'ichi Furuya*
* Oita University, Oita, Japan
E-mail: tomacchi1106@gmail.com

***Abstract*—Automatic chord recognition is a fundamental research topic in music information retrieval (MIR) and has been applied to various tasks, including cover song retrieval and music key detection. In recent years, several methods based on deep learning have been proposed, enabling easier and more accurate recognition. In this study, we focused on the bi-directional Transformer for musical chord recognition (BTC), a chord recognition method based on Transformers, and evaluated its performance. The experimental results showed that the chord recognition accuracy for the triads was 75.52%, which is insufficient for real-world music applications. To improve the recognition accuracy, we focus on the differences in volume between instruments within a song. We propose a method that emphasizes sounds from chord-relevant instruments by applying source separation as a preprocessing step. By amplifying the volume of pitched instruments, such as a piano and guitar, we confirmed that the proposed method successfully corrected more than 2,000 misrecognized chord frames and improved the overall recognition accuracy compared with the conventional method.**

I. Introduction

In music, a chord refers to a harmony consisting of two or more notes played simultaneously, and each chord is given a specific name such as C:maj or D:min depending on the combination of pitches. The sequence of multiple types of chords played in succession is called a chord progression, which serves as a key factor in determining the overall mood of a piece. Identifying the types of chords used in a progression is called chord recognition. Since it enables the understanding of a song's mood and characteristics, it has been utilized in various music-related tasks such as cover song retrieval[1], music key detection, and music structure analysis.[2] However, manual chord annotations by listening to audio require expert knowledge and are time-consuming and costly. Therefore, automatic chord recognition (ACR) is widely adopted. The goal of this study is to improve the accuracy of ACR to enable high-level performance in various music-related tasks.

In recent years, many methods based on deep learning have been proposed that enable chords to be recognized more easily and accurately than with manual annotation. Wu et al. proposed a semi-supervised chord-recognition method using a variational autoencoder that integrates an encoder for recognizing chords and a decoder for generating acoustic features[3], enabling the model to learn from songs without chord labels. Moreover, the methods based on Transformer have also been proposed. Chen et al. introduced a chord-recognition approach for symbolic music, such as MIDI data, using a model [4] that implements a local self-attention mechanism called Intra-Block Intra-MHA and relative positional encoding. Here, MHA represents Multi-Head Self-Attention, which allows self-attention to be performed from multiple perspectives by using multiple heads. In this study, we focused on the bi-directional Transformer for musical chord recognition (BTC), a chord recognition model based on Transformer, and evaluated its performance. The results showed the triads chord recognition accuracy of 75.52%, which is insufficient for applying the model to various music tasks.

To improve recognition accuracy, we focus on the differences in volume between instruments within a song and propose a method that uses source separation as a preprocessing step. Instruments, such as guitars and pianos, which play the component notes of chords, are essential for accurate chord recognition. However, in most contemporary music, unpitched instruments, such as drums and handclaps, and instruments often played with single notes, such as vocals and basses, are frequently used. Consequently, pitched instruments can sometimes be masked by unpitched instruments. In this study, we apply source separation to the original audio to isolate the individual instruments. We then increase the volume of the chord-related accompanying instruments and reconstruct the audio. This allows the model to perform chord recognition using audio in which the harmonic content is emphasized more clearly.

II. Chord Recognition Using BTC

BTC is an ACR model based on the Transformer [5] that employs a bi-directional encoder representation architecture (BERT) [6]. Figure 1 illustrates the overall process of chord recognition using BTC as well as the architecture of BTC.

As shown in Figure 1(a), the BTC model takes a spectrogram obtained by applying the constant-Q transform to raw audio data as the input and outputs the chord labels. As shown in Figure 1(b), the main computational blocks of the BTC model consist of multi-head self-attention and position-wise convolutional blocks. Note that some layers, such as layer normalization and fully connected layers, are omitted from the figure for simplicity. BTC has the advantage of enabling end-to-end training in a single learning phase. In contrast, conventional chord recognition methods, such as conventional neural networks (CNNs) and convolutional recurrent neural networks (CRNNs), often require additional models for tasks

979-8-3315-7206-8/25/$31.00

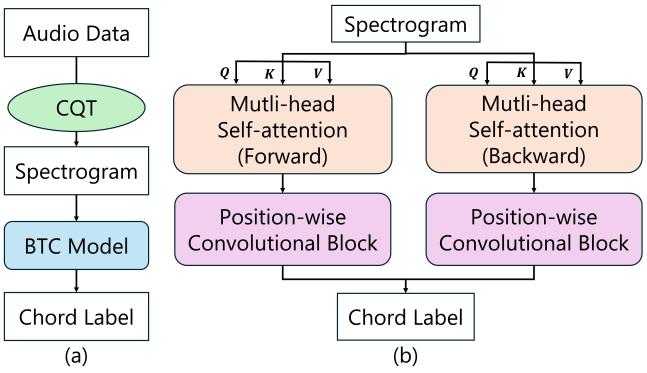

Fig. 1. (a)the process of chord recognition using BTC, (b)the architecture of BTC

such as feature extraction or chord label smoothing. BTC allows for more efficient training than these methods while achieving comparable performance.

An attention mechanism was employed in the multi-head self-attention block. The self-attention mechanism used in BTC is a computational process that examines the relationships between the elements in the input. The output vectors are computed based on queries, keys, and values derived from the input source. Specifically, a weighted sum of the values is calculated using the attention weights derived from the queries and keys, according to the following formula:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Here, Q, K, and V denote the query, key, and value, respectively, and d_k denotes key dimensionality. In the BTC model, bi-directional self-attention is applied, enabling it to learn long-term dependencies from both past and future contexts. Furthermore, by employing a multi-head attention mechanism, the model can perform self-attention from multiple perspectives.

In the Transformer architecture, the features processed by the attention mechanism are nonlinearly transformed on an element-wise basis by a position-wise, fully connected feed-forward network, allowing for richer representations. In the BTC model, instead of a position-wise feed-forward network, a position-wise convolutional block was used. This block replaces the fully connected layers with one-dimensional convolutional layers, enabling the model to incorporate information not only from each individual element but also from the surrounding time frames. By introducing position-wise convolutional blocks, the model could detect chord boundaries across adjacent timeframes and smooth chord sequences. In other words, it helps eliminate unnatural chord transitions (such as single-frame chord anomalies) and enables a smoother representation.

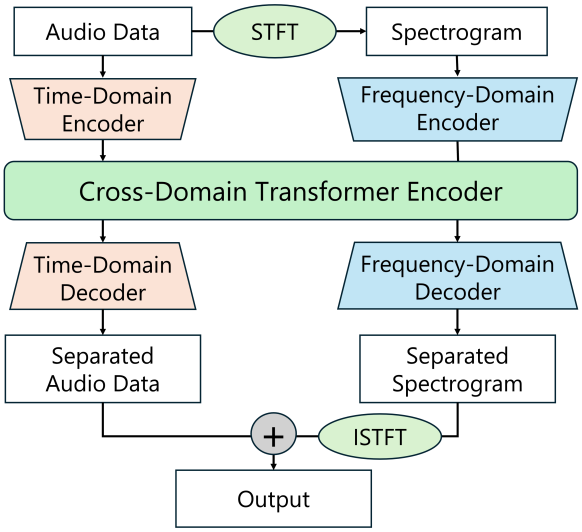

Fig. 2. The process of Music Source Separation using HTDemucs

III. Proposed method

In this study, to improve the accuracy, we focused on the differences in volume between instruments within a song and proposed a method that uses source separation as a preprocessing step.

A. source separation

Source separation is a technique for isolating individual sound sources from an audio signal containing a mixture of various sounds. In this study, we focus on music source separation (MSS) techniques that specifically aim to separate musical sources. MSS aims to separate individual instrument sources from a mixture of multiple instruments, and many methods have been proposed using signal processing techniques such as Harmonic/Percussive Sound Separation (HPSS)[7]. Furthermore, in recent years, many deep learning-based methods have been proposed, including Wave-U-Net [8], which operates in the time domain, and Band-Split RNN [9], which operates in the frequency domain. Methods based on signal processing techniques such as HPSS are easy to implement and capable of separating vocals and percussive instruments. However, it is difficult to finely separate other accompanying instruments into specific sources like bass and guitar. Therefore, in this study, we perform MSS using Hybrid Transformer Demucs (HTDemucs) [10], a deep-learning model based on the Transformers. HTDemucs separates music tracks into four types: vocals, bass, drums, and other, which allows for detailed volume adjustment of each part.

The process of MSS using HTDemucs is illustrated in Figure 2. HTDemucs is a model that combines both time-domain and frequency-domain approaches. HTDemucs adopts a hybrid architecture that processes raw audio in the time domain and spectrograms in the frequency domain. Spectrograms were

obtained by applying a Short-Time Fourier Transform (STFT) to the raw audio. This model uses U-Net-based networks in both domains. The data processed by the encoder in each domain is fed into a cross-domain Transformer encoder, where intra-domain self-attention and inter-domain cross-attention are performed. Afterward, the separated data were output through the decoders in each domain. The data separated in the frequency domain were transformed back to the time domain using the inverse Short-Time Fourier Transform (ISTFT), and then added to the output from the time-domain branch.

B. Volume adjustment

In general, the instruments used in typical music can be broadly classified into two categories: pitched and unpitched. Pitched instruments refer to those whose played notes have a clear pitch, such as guitars and pianos. On the other hand, unpitched instruments are those whose producted sounds are spread over a wide frequency range and do not have a definite pitch, such as drums and hand claps. In most contemporary music, both pitched and unpitched instruments are used; however, their volume levels vary depending on the piece. Consequently, pitched instruments can sometimes be masked by unpitched instruments. However, because chords are determined by combinations of pitches, masking pitched instruments is undesirable. Therefore, in this study, we proposed a method that uses source separation as a preprocessing step to amplify the volume of instruments necessary for chord recognition. While it is possible to completely eliminate unpitched instruments using soruce separation, this would result in the loss of rhythmic information that is important to the song's structure. We believe that by amplifying the volume of pitched instruments while retaining the unpitched instruments, we can emphasize the chord tones while maintaining musical naturalness.

The process flow is illustrated in Figure 3. First, the raw audio data are separated into four instrument sources (vocals, bass, drums, and other) using HTDemucs. Among these, vocals, bass, and other are classified as pitched instruments, so volume adjustment is performed on these sources. The volume adjustment was implemented by manually amplifying the amplitude of the audio's digital signal.

After adjusting the volumes, the separated instrument sounds are recombined into a single audio track, creating a remixed audio in which only the other instruments are emphasized within the original audio. The chord labels can be predicted by feeding this remixed audio into the BTC model.

IV. EXPERIMENT

We conducted experiments to compare the recognition accuracy of the conventional BTC model with that of our proposed method.

A. experimental conditions

For evaluation, 485 songs were used: 225 songs from Isophonics [11], 65 songs from Robbie Williams [12], and 195 songs from uspop2002 [13]. The songs are Western music by artists such as The Beatles and Queen, and consist of genres such as rock, pop, and R&B.

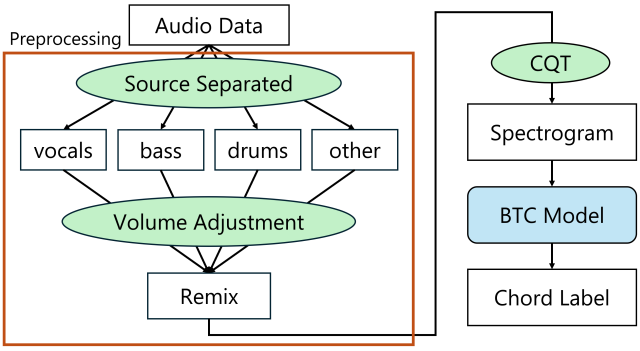

Fig. 3. The process of proposed method

For chord recognition, a publicly available pre-trained model [14] was used. Each audio signal was processed at a sampling rate of 22,050Hz using CQT with six octaves starting from C1, 24 bins per octave, and a hop size of 2048.

Additionally, a large vocabulary was used for the evaluation. The large vocabulary label consists of 170 chord types (12 semitones $\times$ maj, min, dim, aug, min6, maj6, min7, min-maj7, maj7, 7, dim7, hdim7, sus2, sus4 + "X chord" (the unknown chord) + "No Chord"). By mapping each label file to this vocabulary, inconsistencies in chord notation across datasets were eliminated.

The accuracy was evaluated using the Weighted Chord Symbol Recall (WCSR) score calculated over all time frames of each song using the mir_eval library. The formula is as follows, where f_c denotes the number of correctly classified frames, and f_a represents the total number of frames. Additionally, one chord was recognized per frame.

$$WCSR = \frac{f_c}{f_a} \times 100(\%) \tag{2}$$

Using this formula, we compared the triads metrics, which represent the recognition accuracy of triadic chords. For reference, we also compared three other metrics: root, representing root note accuracy; maj-min, representing the accuracy for major and minor chords; and tetrads, representing the accuracy for tetradic chords.

B. results

Table I shows the recognition accuracy of the large vocabulary, comparing the conventional method with the results of amplifying the volume of the vocals, bass, and other parts, respectively. The volume amplification was implemented by doubling the amplitude of each instrument's signal. The table indicates that amplifying the volume of the other achieved the highest recognition accuracy across all metrics.

Table II lists the recognition accuracies when using a large

TABLE I
WCSR SCORES FOR THE LARGE VOCABULARY OF THE CONVENTIONAL METHOD AND THE PROPOSED METHOD WITH MODIFIED VOLUMES FOR VOCALS, BASS, AND OTHER

	triads	root	maj-min	tetrads
Conventional	75.52	82.51	81.59	67.44
vocals	74.75	81.81	80.82	66.48
bass	75.06	82.21	81.09	66.83
other	**75.67**	**82.66**	**81.78**	**67.45**

Unit: %

TABLE II
WCSR SCORES OF LARGE VOCABULARY FOR THE CONVENTIONAL METHOD AND THE PROPOSED METHOD(%)

	triads	root	maj-min	tetrads
Conventional	75.52	82.51	81.59	67.44
Proposed	**75.72**	**82.72**	**81.84**	**67.59**

Unit: %

TABLE III
FRAME-WISE CORRECTNESS EVALUATION OF TRIADS FOR LARGE VOCABULARY BETWEEN THE CONVENTIONAL AND PROPOSED METHOD

		Conventional	
		correct	incorrect
Proposed	correct	847,069	13,675
	incorrect	11,390	264,608

Unit: frames

vocabulary. This result was based on a preliminary experiment in which other instrument volumes were manually adjusted, recognition accuracy was compared, and the best-performing setting was presented. The results in Table II confirm that the proposed method outperformed the conventional method for all metrics.

Furthermore, Table III presents a frame-by-frame comparison of the triad chord labels between the conventional and proposed methods, showing correct and incorrect classifications. A total of 1,136,742 frames were compared. Among these, 11,390 frames were correctly classified using the conventional method, but were misclassified by the proposed method. Conversely, 13,675 frames were misclassified using the conventional method but were correctly classified using the proposed method. From this, it can be said that the proposed method reduced chord classification errors by a total of 2,285 frames in the triads category. We performed a significance test using the sign test on these results and confirmed that the proposed method showed a statistically significant improvement.

C. Discussion

As shown in Table III, the proposed method misclassified 11,390 frames that were correctly identified using the conventional method. This error is considered to be caused by single notes played by instruments such as guitars and pianos. Figure 4 shows the chord label outputs of the conventional and proposed methods, along with the corresponding GroundTruth data. The numbers on the left and center represent the chord

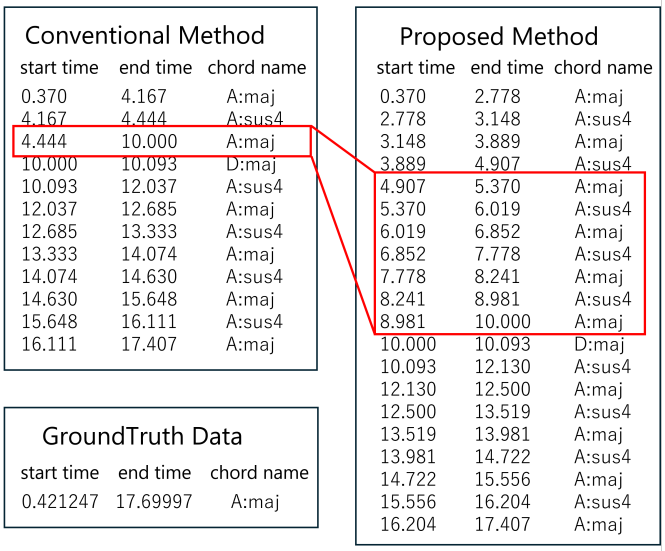

Fig. 4. A part of the chord labels for Doctor Robert

start and end time, respectively, and the right denotes the chord name. According to the GroundTruth data, the A:maj chord continues until 17.7 seconds in this song. However, because a single D note is played in a guitar riff in this section, the conventional method misrecognizes the chord as A:sus4 or D:maj. By amplifying the guitar, which is part of the other instruments, the D note becomes more prominent in the proposed method. This caused the chord label A:maj, which was correctly recognized in the conventional method, to be split into shorter segments, and A:sus4 was often misrecognized instead. This outcome indicates that, while the enhancement of chord tones contributed to improved recognition in many cases, it also amplified isolated notes. These isolated notes, especially when played by the other instruments such as guitars or pianos, sometimes lead to incorrect chord recognition.

V. CONCLUSION

We investigated the performance of BTC which is an automatic chord-recognition model based on the Transformer. To address the issue of insufficient accuracy, we proposed a method that uses source separation as a preprocessing step. We applied music source separation to the original audio data, separating it into four distinct components: vocals, bass, drums, and other. We increased the volume of the other instrument, which was essential for chord recognition, and recombined the instruments into a single audio track. The processed audio was fed into the BTC model to output chord labels. As a result, we confirmed that errors in over 2,000 frames were reduced and the recognition accuracy was improved.

ACKNOWLEDGMENT

Part of this study was supported by JSPS KAKEN 23K28111 and 23H03421.

REFERENCES

[1] K.Lee, "Identifying cover songs from audio using harmonic representation," *MIREX 2006*, pp. 36–38, 2006.

[2] J.Pauwels, F.Kaiser, and G.Peeters, "Combining harmony-based and novelty-based approaches for structual segmentation," in *Proceeding of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil: ISMIR, 2013, pp. 601–606.

[3] Y.Wu, T.Carsault, E.Nakamura, and K.Yoshii, "Semi-supervised neural chord estimation based on a variational autoencoder with latent chord labels and features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2956–2966, 2020.

[4] T.P.Chen and L.Su, "Attend to chords: Improving harmonic analysis of symbolic music using transformer-based models," *Transactions of the International Society for Music Information Retrieval*, vol. 4(1), pp. 1–13, 2021.

[5] A.Vaswani, N.Shazeer, N.Parmar, *et al.*, "Attention is all you need," in *Adcances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[6] J.Devlin, M.W.Chang, K.Lee, and K.Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proceedings of the European Signal Proceeding Conference*, Lausanne Switzerland, 2008.

[8] D.Stoller, S.Ewert, and S.Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[9] Y.Lio and J.Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.

[10] S.Rouard, F.Massa, and A.Défossez, "Hybrid transformers for music source separation," in *Proceeding of 2023 IEEE International Conference on Acoustics Speech and Signal Processing*, Rhodes Island, Greece: ICASSP, 2023.

[11] C.Harte, "Towards automatic extraction of harmony information from music signals," *Ph.D.dissertation, Centre for Digital Music, Queen Mary Univ. London*, 2010.

[12] B. Giorgi, M.Zanoni, A.Sarti, and S.Tubaro, "Automatic chord recognition based on the probabilistic modeling of diatonic modal harmony," in *Proceeding of the 8th International Workshop on Multidimensional Systems*, Erlangen, Germany, 2013.

[13] A.Berenzweig, B.Logan, D. P.W.Ellis, and B.Whitman, "A large-scale evaluation of acoustic and subjective music-similarity measures," *Computer Music Journal*, vol. 28(2), pp. 63–76, 2004.

[14] jayg996, *Btc-ismir19*, https://github.com/jayg996/BTC-ISMIR19, Accessed: 2025-06-09, 2019.