

MIREX 2026 SUBMISSION: AUDIO CHORD ESTIMATION USING THE SHEETSAGE-CHORD DATASET, PRE-TRAINED FEATURES, AND SOURCE SEPARATION

Marco Viviani Xinwei Jiang Zhongyi Huang
Alexandre D’Hooge Yulun Wu Yaolong Ju
Great Bay University Dongguan, Guangdong, China
juyaolong@gbu.edu.cn

ABSTRACT

This abstract describes the system MuQ-BTC, a transformer-based ACE pipeline that departs from classical feature extraction in two aspects: it replaces the Constant-Q Transform (CQT) with embeddings from a frozen self-supervised Music Quantization foundation model (MuQ), and it applies music source separation as a preprocessing step to remove components of the audio mix that do not carry harmonic information (drums and vocals). It is additionally trained on SheetSage-Chord, a new large-scale chord-annotated corpus that we introduce. Experimental results show that our proposed method outperforms existing state-of-the-art approaches in chord estimation performance.

1. INTRODUCTION

Automatic Chord Estimation (ACE) is a Music Information Retrieval (MIR) task that seeks to infer a sequence of chord labels from an audio recording by modeling its harmonic content. Despite recent progress, ACE remains affected by several limitations. First, the datasets commonly employed for training and evaluation are relatively limited in size. Second, many ACE systems rely on the Constant-Q Transform (CQT), a time-frequency representation whose logarithmic frequency resolution is suited to musical pitch and human auditory perception. While self-supervised learning has recently enabled effective feature extraction from large amounts of unlabelled audio in other audio tasks, its potential remains largely unexplored in ACE. Furthermore, audio tracks containing vocals and drums can introduce irrelevant pitches and transient energy, interfering with chord estimation. To address these issues, we introduce SheetSage-Chord, a dataset containing 14148 audio clips with time-aligned chord annotations. We further propose MuQ-BTC, built upon BTC [1], which takes MuQ features as input. Finally, BS-RoFormer [2]

source separation is employed to remove vocals and drums before estimation.

2. METHODOLOGY

2.1 Source separation

The input mixture is processed with BS-RoFormer [2], using the BS-RoFormer-SW six-stem checkpoint. From the six estimated stems (vocals, drums, bass, guitar, piano, other) we discard vocals and drums and reconstruct a chord-oriented mix:

$$x^{\text{mix}} = x^{\text{bass}} + x^{\text{guitar}} + x^{\text{piano}} + x^{\text{other}}, \quad (1)$$

This occurs because the vocals and drums contribute pitch and transient energy components that are largely independent of the underlying harmony.

2.2 Feature extraction

The chord-oriented mix is resampled to 24 kHz and encoded with MuQ [3], a self-supervised music representation model, kept frozen. We take the hidden-state sequence of its final layer: for every 10-second segment this yields 250 frames of 1024-dimensional embeddings, in place of the 144-dimensional CQT features used by the original BTC classifier, described in the following paragraph.

2.3 Classifier

The classifier is the Bidirectional Transformer for Chord recognition [1], left architecturally unchanged. Two hyperparameters were adapted to the new front-end: hidden size raised from 128 to 256 (the smallest capacity that exploited the 1024-dimensional embeddings without underfitting or overfitting in preliminary experiments), and timestep set to $T = 250$ to match MuQ’s 25 Hz frame rate over the 10-second window. Remaining settings follow the original BTC configuration: 8 transformer layers, 4 attention heads, 0.2 as dropout.

2.4 Chord vocabulary

The output alphabet contains 170 classes: 14 chord qualities (*maj*, *min*, *dim*, *aug*, *min6*, *maj6*, *min7*, *minmaj7*, *maj7*, *7*, *dim7*, *hdim7*, *sus2*, *sus4*) over 12 pitch classes, plus the



© F. Author, S. Author, and T. Author. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).
Attribution: F. Author, S. Author, and T. Author, “MIREX 2026 Submission: Audio Chord Estimation Using the SheetSage-Chord Dataset, Pre-trained Features, and Source Separation”, in *Proc. of the 27th Int. Society for Music Information Retrieval Conf.*, Abu Dhabi, UAE, 2026.

no-chord symbol. Enharmonic pitch-class pairs are collapsed into a single class, and the system does not predict bass notes or inversions. Estimated labels are therefore always in root position, and the submission is expected to be penalised on the *majmin-inv* and *sevenths-inv* evaluation vocabularies, where any correctly identified chord in a non-root-position reference will count as a mismatch.

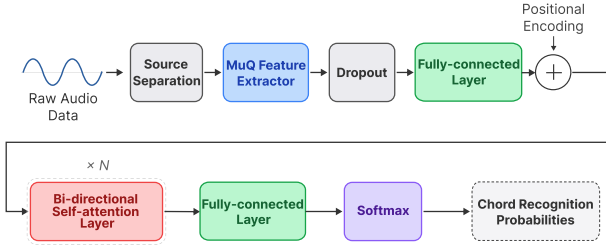


Figure 1. Proposed chord recognition pipeline. Source separation first removes vocals and drums from the input audio. The resulting signal is encoded by the pre-trained MuQ extractor, whose features are processed by the BTC Transformer classifier [1]. The classifier comprises N self-attention blocks with bidirectional layers, positional encoding, and dropout, followed by fully connected and softmax layers that output frame-level chord probabilities.

3. DATASETS

In this section, we report our data usage in full, as required by the task guidelines.

3.1 Existing benchmarks

We used some of the most cited datasets in the literature for ACE tasks: Isophonics [4] (174 songs by The Beatles), RWC-Pop [5] (100 songs), and Robbie Williams [6] (59 songs). These constitute our primary annotated training corpus. We additionally used 15 Japanese tracks [7] as an external test set, excluded from the training and validation phases to test the model on previously unseen data.

3.2 SheetSage-Chord

A new dataset of 14148 audio clips (not longer than 1 minute approximately) for a total of 88 hours, built by us from the symbolic annotations of the SheetSage release [8] and paired with the corresponding audio recordings. The resulting 14148 clips originate from 8866 unique songs by 4107 artists and span 32 genres. The dataset will be released as chord labels with timestamps under CC BY-NC-SA 3.0, inherited from the original SheetSage annotations, together with SheetSage IDs for audio retrieval and pre-computed audio features. Raw audio is not redistributed, for copyright reasons. No split of Billboard 2013 was used at any stage of development: our SheetSage-Chord dataset consists of audio clips of at most one minute in duration, and its chord annotations follow a convention distinct from that of Billboard.

Table 1. Dataset splits used in this work. Isophonics, RWC-Pop, Robbie Williams and SheetSage-Chord are used for training and validation, Idol-Songs-JP for testing only. All values denote numbers of songs for each model of the 5-fold cross-validation: the five folds share these split sizes but partition the songs differently.

Dataset	Train	Val	Test
Isophonics [4]	139	35	–
RWC-Pop [5]	80	20	–
Robbie Williams [6]	47	12	–
Idol-Songs-JP [7]	–	–	15
SheetSage-Chord (Proposed)	11318	2830	–

5-fold cross-validation is applied to calculate scores for all datasets except Idol-Songs-JP: five different models are trained using a split of 80% for training and 20% for validation.

4. EXPERIMENTS

4.1 Data augmentation

Offline preprocessing pitch-shift augmentation was applied to the training split of every dataset, with $k \in \{-5, -4, -3, -2, -1, 0, +1, +2, +3, +4, +5, +6\}$ semitones, covering all twelve pitch classes. Both waveform and chord labels are transposed by the same amount. Pitch-shifting uses the Rubber Band Library’s ¹ algorithm. No augmentation is applied to validation or test splits.

4.2 Optimization

We applied frame-level cross-entropy, Adam (learning rate of 10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$), decay rate of 0.95 applied when validation accuracy fails to improve, and batch size of 64. Input features are standardised per fold using statistics computed on that fold’s training partition. Training is capped at 100 epochs with early stopping after 10 epochs without improvement in validation accuracy; the checkpoint with the best validation accuracy is retained.

4.3 Submitted checkpoint

The models are trained using 5-fold cross-validation on Isophonics, RWC, Robbie Williams, and SheetSage-Chord, meaning that each checkpoint has seen only 80% of that data. For MIREX 2026, a single model is submitted, specifically the one from fold 0 (the best-performing one).

5. RESULTS

We report results for the state-of-the-art open-source ACE systems available at the time of writing, BTC [1] and CNN-BiLSTM [9], which serve as our reference baselines

¹ <https://breakfastquay.com/rubberband/>

Table 2. Macro-average scores on the four evaluation datasets of **IRRI** (Isophonics, RWC, Robbie Williams, the external Idol-Songs-JP), and on the new SheetSage-Chord dataset. Six model configurations are presented: [1] and [2] serve as baselines, and [3] to [6] are derived via successive incremental additions based on the prior configuration each time, where [6] represents our proposed MuQ-BTC model. The best result per column is shown in **bold**, the second-best is underlined.

Configuration	Root	Thirds	Triads	Sevenths	Tetrads	Maj-min	MIREX
<i>IRRI</i>							
[1] CNN-BiLSTM	82.77%	78.58%	76.25%	68.08%	63.86%	80.23%	78.05%
[2] BTC baseline	83.70%	80.52%	77.62%	68.98%	64.36%	81.78%	80.71%
[3] + MuQ (in place of CQT)	85.57%	81.52%	78.58%	70.10%	65.30%	83.05%	82.18%
[4] + SheetSage-Chord	86.14%	82.78%	79.96%	72.18%	67.44%	84.00%	83.54%
[5] + SheetSage-Chord augmented	<u>86.85%</u>	<u>83.61%</u>	<u>81.05%</u>	<u>73.34%</u>	<u>68.77%</u>	<u>84.79%</u>	<u>84.38%</u>
[6] + Source Separation (the proposed MuQ-BTC)	87.21%	84.31%	81.59%	73.72%	69.10%	85.43%	84.93%
<i>SheetSage-Chord</i>							
[4]	77.57%	74.08%	72.14%	64.64%	61.97%	75.32%	74.88%
[5]	78.96%	75.94%	74.01%	66.82%	64.28%	77.03%	76.65%
[6]	<u>78.51%</u>	<u>75.15%</u>	<u>73.23%</u>	<u>66.21%</u>	<u>63.70%</u>	<u>76.21%</u>	<u>75.85%</u>

and retrained with our datasets. To isolate the contribution of each design choice in our pipeline, we then conduct an ablation study on BTC, progressively adding components until reaching the MuQ-BTC configuration proposed in this work. Table 2 reports the ablation of our work, as macro-averaged Weighted Chord Symbol Recall (WCSR) over the four evaluation sets IRRI (Isophonics, RWC, Robbie Williams, Idol-Songs-JP), with five-fold cross-validation. "MIREX" denotes the `mir_eval` MIREX metric (at least three shared pitch classes).

Both baselines were retrained from scratch on the same IRRI corpus rather than reusing published weights. The full system improves over the BTC baseline by 4.22 MIREX points; the largest single contributions come from the MuQ front-end (+1.47) and from the addition of augmented SheetSage-Chord data (+2.20), with source separation contributing a further +0.55.

6. CONCLUSIONS

MuQ-BTC combines a frozen self-supervised music encoder, source separation preprocessing, and a bidirectional transformer classifier, trained on a corpus enlarged by the new SheetSage-Chord dataset. Overall, our work illustrates how this approach can move ACE beyond the plateau observed in the past years. Extending MuQ-BTC to predict bass notes and inversions is the natural next step.

7. REFERENCES

- [1] J. Park, K. Choi, S. Jeon, D. Kim, and J. Park, "A bi-directional transformer for musical chord recognition," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*, Delft, The Netherlands, 2019, pp. 620–627. [Online]. Available: <http://archives.ismir.net/ismir2019/paper/000075.pdf>
- [2] W. T. Lu, J. C. Wang, Q. Kong, and Y. N. Hung, "Music source separation with band-split rope transformer," in *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024)*. Seoul, South Korea: IEEE, 2024, pp. 481–485.
- [3] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, "Muq: Self-supervised music representation learning with mel residual vector quantization," 2025. [Online]. Available: <https://arxiv.org/abs/2501.01108>
- [4] M. Mauch, "Automatic chord transcription from audio using computational models of musical context," Ph.D. dissertation, Queen Mary University of London, London, UK, 2010.
- [5] 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 2002)*, Paris, France, 2002, pp. 287–288.
- [6] 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 2013)*, Erlangen, Germany, 2013.
- [7] H. Suda, J. Koguchi, S. Yoshida, T. Nakamura, S. Fukayama, and J. Ogata, "Idolsongsjp corpus: A multi-singer song corpus in the style of japanese idol groups," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR 2025)*, Daejeon, South Korea, 2025, pp. 647–654.
- [8] C. Donahue, J. Thickstun, and P. Liang, "Melody transcription via generative pre-training," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.
- [9] J. Jiang, K. Chen, W. Li, and G. Xia, "Large-vocabulary chord transcription via chord structure decomposition," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*, Delft, The Netherlands, 2019, pp. 644–651. [Online]. Available: <http://archives.ismir.net/ismir2019/paper/000078.pdf>