FROM DISCORD TO HARMONY: DECOMPOSED CONSONANCE-BASED TRAINING FOR IMPROVED AUDIO CHORD ESTIMATION

Andrea Poltronieri **Xavier Serra** **Martín Rocamora**
Music Technology Group, Universitat Pompeu Fabra
{andrea.poltronieri, xavier.serra, martin.rocamora}@upf.edu

ABSTRACT

Audio Chord Estimation (ACE) holds a pivotal role in music information research, having garnered attention for over two decades due to its relevance for music transcription and analysis. Despite notable advancements, challenges persist in the task, particularly concerning unique characteristics of harmonic content, which have resulted in existing systems' performances reaching a glass ceiling. These challenges include annotator subjectivity, where varying interpretations among annotators lead to inconsistencies, and class imbalance within chord datasets, where certain chord classes are over-represented compared to others, posing difficulties in model training and evaluation. As a first contribution, this paper presents an evaluation of inter-annotator agreement in chord annotations, using metrics that extend beyond traditional binary measures. In addition, we propose a consonance-informed distance metric that reflects the perceptual similarity between harmonic annotations. Our analysis suggests that consonance-based distance metrics more effectively capture musically meaningful agreement between annotations. Expanding on these findings, we introduce a novel ACE conformer-based model that integrates consonance concepts into the model through consonance-based label smoothing. The proposed model also addresses class imbalance by separately estimating root, bass, and all note activations, enabling the reconstruction of chord labels from decomposed outputs.

1. INTRODUCTION

In Western music theory, chords denote simultaneous combinations of three or more notes, forming harmonic structures integral to musical composition and analysis [1–4]. However, manually annotating chords from audio recordings is a labour-intensive task requiring music professionals' expertise. Consequently, Audio Chord Estimation (ACE) emerged as a crucial task in Music Information Retrieval/Research (MIR) to automate chord transcription from audio due to its relevance for its numerous applications in music transcription and analysis.

The research in ACE has witnessed more than two decades of exploration, but despite the important advancements achieved [5], performance results have stagnated in recent years, leading some researchers to suggest that the task has hit a glass ceiling [6]. These challenges stem from several significant open problems [5], which are fundamentally linked to the complex nature of harmonic content and its representation within audio signals.

One such challenge is the chord vocabulary imbalance, stemming from the unequal frequency of occurrence among chord labels. For instance, in *ChoCo* [7], the most extensive corpus of chord annotations to date, approximately 74.9% of the distribution of the 8064 distinct chord classes is dominated by just major, minor, major seventh, minor seventh and dominant seventh chord types.

Another critical challenge is inter-annotator agreement, which arises from the inherent ambiguity in what constitutes a chord from a musical perspective and the subjective nature of human annotation processes. For example, a clear distinction between a chord sequence and a melodic line can be subject to individual interpretation. Moreover, there is significant variance among annotators regarding the level of detail in annotating chord sequences [8].

Various studies have investigated inter-annotator agreement in chord annotation, reporting agreement rates for the root note ranging from 76% [8] to 92% [9], using different datasets and numbers of annotators. Such evaluations typically use binary metrics to compare labels, but penalising the agreement evaluation equally for every discrepancy can be inappropriate [10]. Indeed, binary evaluation risks overlooking harmonic aspects that might be shared among chord sequences, although annotated differently.

As a preliminary contribution of this paper, we analyse patterns of inter-annotator disagreement in chord annotation. Our analysis reveals that when annotators disagree, their chord labels tend to be harmonically related rather than randomly different. Specifically, we find that disagreements commonly occur between chords that share significant harmonic content (c.f. Section 3.1).

Building upon these insights, we propose a method for incorporating such information into the supervised training of ACE systems. Hence, we introduce a novel model integrating consonance-based label smoothing [11] (c.f. Section 3.2.2). To tackle the class imbalance issue, instead of mapping audio features to a predetermined vocabulary of chord labels, we adopt an approach inspired by [12], in which the chord root, bass, and all note activations are classified separately. The final predicted chord label is derived from decoding these three sets of information without explicitly imposing any vocabulary on it (c.f. Section 3.2.1).

The proposed model leverages the Conformer architecture [13], which has recently been explored in several music audio applications [14–16]. We demonstrate that the proposed model performs better than the state-of-the-art approaches, especially when evaluated using non-binary and consonance-based distance metrics (c.f. Section 4).

2. RELATED WORK

Since Fujishima's early work [17], chord recognition has followed knowledge-driven approaches [18], typically extracting chroma [19] or Tonnetz features [20], and classifying them via HMMs, DBNs [19], or CRFs [21].

With the emergence of deep learning, various architectures have been explored for the task, including Convolutional Neural Networks (CNNs) [12, 21], Recurrent architectures (RNN) [22], Convolutional Recurrent Neural Networks (CRNNs) [23], and Transformers [24]. While deep-learning approaches have surpassed traditional knowledge-driven ones, several challenges must be tackled. Most of the proposed approaches to addressing the chord class imbalance challenge can be divided into two categories: chord simplification and chord decomposition. The former reduces the size of the chord vocabulary by converting complex chord labels into simpler representations. Notably, the vast majority of studies have adopted restricted vocabularies of approximately 25 symbols, encompassing major-minor chords [17, 18]. Chord decomposition strategies focus on predicting the chord constituting components separately, and then map them to templates to predict the final chord [12,23,25]. Some additional approaches do not fall into these two categories, like addressing the unequal distribution of chords through a balanced learning process [26], or using a curriculum learning training scheme to begin with simple chord qualities and then move to more complex and less common ones [27].

The inter-annotator agreement in chord annotation continues to pose a significant challenge. Despite existing diagnoses and quantification of this phenomenon in the literature [8, 9], definitive solutions have yet to emerge. Clercq et al. [9] observe an inter-annotator agreement rate of 94% for the root note between two different annotations of the top 20 tracks from Rolling Stone magazine's list of the *500 Greatest Songs of All Time*. In contrast, Koops et al. [8] report an inter-annotator agreement rate of 76% for the root note on four different annotations of a 50-song subset of the Billboard dataset [28]. To address annotation subjectivity, Koops et al. [8, 29] propose a personalised chord estimation framework that adapts labels to individual annotator vocabularies. Their method computes Shared Harmonic Interval Profiles (SHIPs) from multiple reference annotations aligned with CQT frames and trains a neural network to predict user-specific chord labels, offering an alternative to fixed-vocabulary systems. While this approach offers valuable insights into annotation variability, it addresses personalization rather than resolving fundamental inter-annotator disagreement. In contrast, our proposed method develops generalized harmonic representations grounded in music theory principles, thereby eliminating dependence on predefined chord vocabularies.

Moreover, our method applies Label Smoothing (LS), a technique employed to enhance the generalisation and learning speed of multi-class neural networks. Originally proposed in [30], LS redistributes a portion of the probability mass from the observed class to other classes, thereby softening the distribution and generating what is referred to as *soft targets*. This regularisation method has found widespread application in various state-of-the-art models across domains such as image classification, language translation, and speech recognition. It has also been tested for music classification tasks [31], improving performance and reducing overfitting in small network training.

While LS primarily serves as a regularisation technique, numerous studies have delved into its potential for encoding meaningful relationships among different categories. For instance, in [32], authors propose an impactful method for generating more reliable soft labels that explicitly consider the relationships among various categories. Similarly, in [33], a novel approach known as *label relaxation* is introduced, which involves replacing a degenerate probability distribution associated with an observed class label, not by a single smoothed distribution but rather by a larger set of candidate distributions.

We integrate label smoothing into a model based on the conformer architecture [13], which has recently emerged in Automatic Speech Recognition (ASR) as an effective way of modelling global and local audio dependencies by leveraging a combination of CNNs and Transformer architectures. It has showcased remarkable success across various tasks not only in speech [34] but also in music [15], including melodic transcription [14], representation learning [35], and music audio enhancement [36]. It also proved to be suitable for harmonic analysis, as it has been used for audio–chord alignment [16] and more recently adapted for chord estimation [37], where it is combined with the large-vocabulary decoding scheme proposed in [23].

3. METHODS

We present a four-part investigation into chord estimation:

(i) we conduct a comprehensive analysis of inter-annotator agreement across multiple chord similarity metrics, assessing how non-binary metrics measure inter-annotator agreement scores;

(ii) we introduce a new perceptually-informed distance metrics and we demonstrate how it can improve agreement between annotators;

(iii) we introduce a consonance-based label smoothing that leverages consonance to improve chord recognition;

(iv) we present a novel chord label encoding/decoding methodology, inspired by [12].

3.1 Analysis of Inter-Annotator Agreement

As outlined in Section 1, standard metrics employed to evaluate chord estimation systems have traditionally relied on binary comparison approaches [5]. The most fundamental of these is the binary distance $B_{\text{dist}}(C_1, C_2)$, which is defined as 1 if $C_1 = C_2$, and 0 otherwise.

When evaluating chord annotations or estimation algorithms, this binary comparison is typically weighted by the duration of each chord segment to compute the Chord Symbol Recall (CSR) [38]:

$$CSR = \frac{\mid S_a \cap S_e \mid}{\mid S_a \mid}. \tag{1}$$

where S_e represents the set of time segments where the estimated chords match the reference annotations, and S_a represents the total duration of annotated segments.

In addition to overall binary agreement, several granular evaluation metrics have been introduced, each capturing different levels of harmonic detail. The *Root* metric compares only the root note, ignoring chord quality and extensions. *Thirds* extends this by incorporating major and minor third intervals. *Triads* evaluate the full triadic structure—including major, minor, augmented, diminished, and suspended chords, up to the fifth scale degree. *Tetrads* consider closed-voicing chords with extended tones (e.g., 9ths, 11ths, 13ths) collapsed into a single octave. The *Sevenths* metric restricts evaluation to a predefined set of common seventh chord types. Finally, the *MIREX* metric deems an estimate correct if it shares at least three pitch classes with the reference chord, regardless of root or quality. These metrics can optionally account for chord inversions by requiring the bass note to match as well. All are implemented in the `mir_eval` library [39], which is the de facto standard for chord estimation evaluation. These metrics have been consistently used in literature to assess inter-annotator agreement in chord datasets, reporting agreement rates for the root note ranging from 76% [8] to 92% [9].

However, to overcome the inherent limitations of binary evaluation metrics, recent research has introduced alternative measures. McLeod et al. [10] proposed three new metrics that more accurately represent musical relationships among chords: Spectral Pitch Similarity, Tone-by-Tone Distance, and Mechanical Distance.

Spectral Pitch Similarity, which assesses perceived pitch content based on psychoacoustic principles, lies beyond the scope of this study. On the other hand, *Tone-by-Tone Distance* (TbT) treats chords as pitch-class sets, categorising pitches as either tonal or neutral. This metric quantifies chord similarity by measuring the proportion of shared pitch classes, resulting in a distance value reflecting their pitch-content similarity. In contrast, *Mechanical Distance* provides a more granular evaluation by approximating the physical distance between chord labels as they would be played on an instrument. It extends Tone-by-Tone Distance by quantifying not only the proportion of incorrect pitches but also the magnitude of each deviation from the target chord, by default measured in semitones.

While this approach introduces a more musically grounded notion of distance, the original formulation of Mechanical Distance still treats all semitone deviations as perceptually equivalent. This simplification overlooks the fact that, in Western tonal harmony, the perceptual impact of an interval depends not only on its size but also on its harmonic function. To address this limitation, we propose an extension that incorporates consonance-based weighting into the Mechanical Distance. Specifically, we introduce the Mechanical-Consonance metric, which integrates the perceptual consonance vector presented in [40], grounded in empirical studies of Western tonal harmony.

The consonance vector is defined as:

$$vt = [0, 7, 5, 1, 1, 2, 3, 1, 2, 2, 4, 6] \tag{2}$$

where each position corresponds to an interval in semitones, assigning lower values to more consonant intervals. For instance, perfect fifths and thirds (P5, m3, M3) receive the lowest score (1), indicating high consonance, while dissonant intervals such as major sevenths, minor seconds, and tritones are assigned higher values (up to 7). Intervals of intermediate consonance, such as fourths and sixths, are assigned moderate values. By weighting semitone deviations using this vector, the Mechanical-Consonance metric adjusts the contribution of each error based on its perceptual salience.

As a first contribution of this paper, we assess inter-annotator agreement across various chord granularity levels (e.g., root, thirds, triads) by comparing standard `mir_eval` metrics with Tone-by-Tone Distance and Mechanical Distance. To align these non-binary metrics with the granularity levels typically employed in ACE evalu-

	CASD Dataset			
Metric	**mir_eval↑**	**TbT↑**	**Mech↓**	**Mech-Cons↓**
Root	0.757	0.773	0.817	0.604
Thirds	0.741	0.773	0.896	0.716
Triads	0.710	0.796	1.549	1.663
MajMin	0.734	0.803	1.465	1.577
Tetrads	0.572	0.786	1.859	1.803
Sevenths	0.592	0.794	1.771	1.715
MIREX	0.744	0.786	1.859	1.803
	Random Dataset			
Metric	**mir_eval↑**	**TbT↑**	**Mech↓**	**Mech-Cons↓**
Root	0.145	0.158	2.914	2.336
Thirds	0.140	0.158	2.914	2.336
Triads	0.121	0.253	5.536	5.861
MajMin	0.124	0.248	5.530	5.958
Tetrads	0.121	0.253	5.536	5.861
Sevenths	0.124	0.248	5.530	5.961
MIREX	0.121	0.253	5.536	5.861

Table 1. Inter-Annotator Agreement Scores for Chord Annotations. TbT = Tone-by-Tone distance, Mech = Mechanical distance, Mech-Cons = Mechanical with Consonance distance.

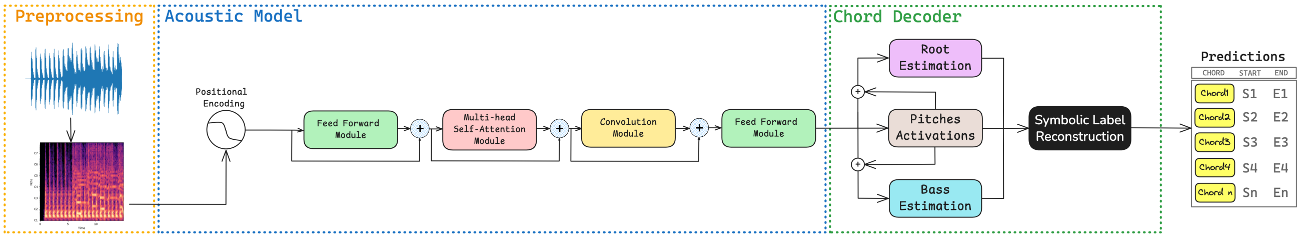

Figure 1. Overview of the Conformer model architecture, which comprises the preprocessing stage, the conformer-based model, and the symbolic chord decoder.

ations, we apply two heuristics: (i) restricting comparisons to the pitch ranges considered by the respective `mir_eval` metrics (e.g., pitches up to the fifth of the chord for the *MajMin* metric); and (ii) limiting comparisons only to chords included in the `mir_eval` metric evaluation (e.g., diminished and seventh chords are excluded from the *MajMin* metric).

We conduct this analysis on the Chordify Annotator Subjectivity Dataset (CASD) [41], which represents the largest available dataset for assessing chord annotation agreement and was previously used for similar studies [8].

Moreover, to establish baseline performance and assess metric reliability, we conduct parallel experiments on a synthetically generated dataset replicating CASD's structure (50 tracks with 4 annotations each), but populated with randomly generated chord sequences that preserve both its chord vocabulary and sequence-length distributions.

Table 1 reports the results for both the CASD and synthetic datasets, highlighting the performance and reliability of each metric across different evaluation settings. To aid interpretation, we first clarify the nature and scaling of each metric under comparison.

The `mir_eval` metrics are formulated as similarity measures, returning values in the range $[0, 1]$, where 1 indicates perfect agreement and 0 indicates complete disagreement. In contrast, Tone-by-Tone Distance is defined as a distance metric in $[0, 1]$, with 0 indicating identical pitch-class content and 1 indicating no overlap; we convert it to a similarity score by computing $1 - \texttt{TbT}$. Mechanical Distance returns an unbounded distance value influenced by the number of notes in the chords, the sequence length, and the underlying pitch distance function. Due to these variable factors, we report Mechanical Distance in its original form without normalisation, as any fixed rescaling would obscure meaningful differences.

The results show a clear separation between the CASD and random datasets, confirming that all metrics are sensitive to musically meaningful agreement. TbT similarity scores are remarkably stable across all chord granularity levels, including more complex ones such as Sevenths and Tetrads. In the random dataset, TbT returns consistently higher values than `mir_eval`, and scores increase progressively as more notes are considered in the evaluation (e.g., from Root to Sevenths). This trend indicates that TbT is more permissive than discrete match-based approaches and more sensitive to coincidental pitch-class overlap when more components are involved.

Mechanical Distance exhibits lower agreement for simpler structures (e.g., Root and Thirds), closely mirroring the `mir_eval` pattern. This is also reflected in the random dataset, where increasing the chord complexity leads to proportionally larger distances.

Mechanical-Consonance generally produces lower scores for the CASD dataset and higher scores for the random dataset compared to its unweighted counterpart. Notably, the mean difference between CASD and random results is 3.326 for Mechanical Distance and 3.471 for Mechanical-Consonance. This larger separation supports the idea that inter-annotator disagreements are not random but often occur between harmonically related chords. The consonance-weighted formulation reinforces this insight by penalising perceptually dissonant deviations more heavily, further distinguishing musically plausible disagreements from unstructured noise.

3.2 Proposed Model

As a second contribution, this paper presents a novel ACE model, illustrated in Figure 1, which leverages the Conformer architecture [13]. As a first step, the audio is first resampled to a sampling rate of 22050 Hz, and a hop size of 2048 is applied. Then, the Constant-Q Transform (CQT) features are calculated on 6 octaves starting from $C1$, with 24 bins per octave, resulting in a total of 144 bins. The CQT features are fed to a conformer encoder [13] before being passed to the decoder layers.

3.2.1 Chord Decomposition and Decoding

Label encoding follows a similar approach as [12]. Root and bass notes are encoded as a 13-dimensional one-hot vector, where the first 12 positions represent the semitones from C to B, and the last one indicates silence (denoted as N). Chord tones are encoded using a 12-dimensional multi-hot vector, where each dimension indicates the presence (1) or absence (0) of a pitch class in the chord.

The output of the Conformer layers is first passed through a fully connected head to predict chord tones. These chord predictions then serve as conditioning information for two additional components: bass and root prediction. Each of these components employs a feature fusion mechanism that concatenates the original Conformer features with the chord logits, creating an enriched representation that captures both the acoustic context and the predicted harmonic content. This hierarchical approach re-

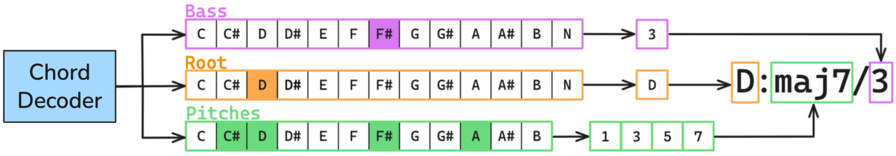
Chord Decoder
Bass
C C# D D# E F F# G G# A A# B N
3
Root
C C# D D# E F F# G G# A A# B N
D
Pitches
C C# D D# E F F# G G# A A# B
1 3 5 7
D:maj7/3

Figure 2. Example of chord label decoding for a `D:maj7/3` chord using the decomposed decoder, inspired by [12].

flects the musical intuition that bass and root notes are contextually dependent on the overall harmonic content, rather than treating all three components as independent prediction tasks. To train the model, we use a composite loss that aligns with this encoding scheme. Cross-entropy loss is applied to root and bass predictions, and binary cross-entropy loss is used for chord tone predictions. Additionally, we introduce a regularisation term that penalises discrepancies between the predicted and actual number of active pitch classes.

The total loss is defined as:

$$\mathcal{L} = \lambda_{\text{root}}\, \mathcal{L}_{\text{CE}}^{\text{root}} + \lambda_{\text{bass}}\, \mathcal{L}_{\text{CE}}^{\text{bass}} + \lambda_{\text{chord}}\, \mathcal{L}_{\text{BCE}}^{\text{chord}} \\ + \lambda_{\text{card}}\, \|\hat{c} - c\|_1 \tag{3}$$

where c and $\hat{c}$ are the number of active notes in the ground truth and those predicted above a threshold, respectively.

Differently from [12], where the outputs of the bass, root, and pitch activation predictions are combined and passed through a final linear layer to predict chord labels, we directly use these three components to reconstruct the final chord label. The novelty of this approach lies in the fact that, unlike vocabulary-constrained decoding strategies such as [23], our method does not require a predefined chord vocabulary.

Chord labels are reconstructed from the predicted probabilities in a modular decoding process. First, the root note is identified by selecting the pitch class with the highest predicted probability, which is then mapped to its symbolic representation. For the chord tones, a fixed threshold (default: 0.5) is applied to the predicted pitch activations; only pitches exceeding this threshold are retained. These pitch classes are then converted into intervals relative to the predicted root. An analogous procedure is applied to the bass prediction, allowing the full reconstruction of the chord structure, as illustrated in Figure 2. Finally, the decoded chord is passed to the `harte_library`[1], which implements utilities for converting the predicted chord label into the respective shorthand notation.

3.2.2 *Consonance-based Smoothing*

We introduce a novel label smoothing technique that leverages music-perceptual knowledge by incorporating consonance relationships between pitch classes. Unlike conventional label smoothing that uniformly distributes probability mass across incorrect classes, our approach allocates probability according to the consonance relationship between pitch classes.

Let $\mathbf{c} = [c_0, c_1, \ldots, c_{11}] \in \mathbb{R}^{12}$ be a consonance vector where each element c_i quantifies the dissonance level of the interval i semitones above the reference pitch. Lower values of c_i indicate more consonant intervals (e.g., perfect fifth, major third). We transform this vector into a similarity measure $\mathbf{s} \in \mathbb{R}^{12}$ as follows:

$$\mathbf{s} = 1 - \frac{\mathbf{c}}{\max(\mathbf{c})} \tag{4}$$

This ensures that more consonant intervals receive higher similarity scores, with perfect consonance (unison) having a similarity of 1. For a given target pitch class $t \in \{0, 1, \ldots, 11\}$ and smoothing factor $\alpha \in [0, 1]$, we define the smoothed target distribution $\mathbf{q} \in \mathbb{R}^{12}$ as:

$$q_i = \begin{cases} 1 - \alpha & \text{if } i = t \\ \alpha \cdot s_{(i-t) \bmod 12} & \text{if } i \neq t \end{cases} \tag{5}$$

The distribution is then normalised to ensure $\sum_{i=0}^{11} q_i = 1$:

$$\mathbf{q} = \frac{\mathbf{q}}{\sum_{i=0}^{11} q_i} \tag{6}$$

This formulation creates a probability distribution where the target class t receives the highest probability $(1 - \alpha)$, while the remaining probability mass α is distributed among other pitch classes proportionally to their consonance relationship with the target. For example, when the true class is C (0), pitch classes G (7) and F (5) will receive higher probability than more dissonant intervals like C# (1) or B (11), reflecting their stronger harmonic relationships.

4. EVALUATION

In this section, compare the performance of the proposed ACE model with a state-of-the-art method [24], using standard `mir_eval` metrics, Tone-by-Tone (TbT) similarity, and Mechanical distances. Additionally, we evaluate the effectiveness of the proposed chord decoder by benchmarking it against a conventional frame-wise classification approach, focusing on its ability to accurately capture chord inversions using the inverted `mir_eval` metrics.

All chord annotations were sourced from ChoCo [7], which provides standardized labels in Harte syntax [42]. Specifically, we use annotations from the Isophonics dataset [43] and the McGill Billboard corpus [44] for training and validation, while the RWC Pop [45] and USPop datasets [23] serve as test sets. This setup enables evaluation of both model performance and generalization across diverse chord vocabularies.

[1] `https://github.com/andreamust/harte-library`

Model	Vocab	Smooth	Root↑	MajMin↑	Thirds↑	Triads↑	Tetrads↑	7th↑	MIREX↑	TbT↑	Mech↓	MechCons↓
Ours	170	-	81.4	77.5	78.1	72.3	59.6	64.7	79.4	77.9	1.55	1.35
Ours	Decom.	-	83.4	77.2	79.7	72.2	59.2	64.6	79.3	80.5	1.57	1.37
Ours	Decom.	Cons.	**84.0**	**77.8**	**80.3**	**72.7**	**60.8**	**66.0**	**79.8**	**81.7**	**1.44**	**1.30**
BTC	170	-	81.6	77.3	78.4	72.1	60.0	65.7	79.0	78.4	1.60	1.40
BTC	Decom.	-	82.9	76.0	79.2	70.9	57.2	62.4	77.4	80.4	1.52	1.35
BTC	Decom.	Cons.	82.8	76.1	79.3	70.9	59.5	64.7	79.0	80.7	1.49	1.32

Table 2. Performance comparison across different model variants using both standard `mir_eval` metrics and non-binary metrics. Results are reported for our conformer-based model with and without the decomposition decoder and consonance-based label smoothing. Additionally, we compare these settings with the BTC model [24].

To increase data density while preserving local harmonic continuity, each track is segmented into 20-second excerpts with 50% overlap. We employ data augmentation by transposing both audio and targets from -5 to $+6$ semitones. During training, we use the *AdamW* optimiser and cosine annealing learning rate scheduler to dynamically adjust the learning rate during training cycles. Additionally, we adopted mixed precision training [46] to accelerate training. To prevent overfitting, we implement early stopping, terminating training when performance on a validation set ceased to improve after 10 epochs. The code and all hyper-parameters used in the experiments are available on the GitHub repository of the project [2] .

Metric	BTC	Ours	Ours	Ours
Vocab.	170	170	Decom.	Decom. Cons.
MajMin Inv.↑	71.5	72.4	**75.6**	**75.6**
Thirds Inv.↑	72.6	72.9	77.2	**77.9**
Triads Inv.↑	67.2	67.6	70.2	**70.8**
Tetrads Inv.↑	56.2	55.7	57.7	**59.4**
Sevenths Inv.↑	60.8	60.0	62.9	**64.4**

Table 3. Performance comparison on inverted chords between traditional architectures and the proposed decomposed model, evaluated using `mir_eval` metrics.

4.1 Evaluation of the ACE Model

We evaluate our model using TbT similarity, Mechanical Distance, and its consonance-weighted variant, as introduced in Section 3.1, alongside standard binary metrics from `mir_eval` [39]. For comparison, we adopt the BTC model [24], a state-of-the-art baseline for audio chord estimation. We reimplemented and retrained the BTC model using the hyper-parameter settings specified in the original paper, enabling a direct evaluation of our proposed decomposition-based decoder and the impact of consonance-informed label smoothing. The experimental results are summarized in Table 2.

As noted by [23], differences among models are often marginal when evaluated with standard metrics. This holds true in our comparison: both models yield similar results on the standard classification task over a 170-class chord vocabulary. However, our proposed decomposition-based decoder consistently outperforms the standard frame-wise classification architecture across several metrics, with the advantage of not relying on a fixed chord vocabulary. Notably, we observe the greatest improvement in Root and Thirds metrics. Additionally, the use of non-binary metrics further highlights the benefits of the proposed decoder. As shown in Table 3, inverted metrics also improve when using the proposed decomposed decoder. This improvement stems from the fact that the proposed chord decoding scheme explicitly predicts the bass note, enabling accurate inversion prediction–a capability that standard chord classification approaches inherently lack. The same trend is confirmed when applying the decomposed decoder to the BTC model, which yields performance increases across several metrics, especially the non-binary ones.

When integrating the consonance-weighted loss on root and bass predictions within the proposed decoding architecture, performance slight improvement on all metrics. Notably, improvements are observed also on non-binary metrics and on inverted chords. The trend is also confirmed when applying consonance smoothing to the BTC model with the decomposed decoder. Overall, evaluation results suggest that both the proposed decoder and the consonance smoothing improve accuracy in most metrics, and led to predictions more consonant to the target.

5. CONCLUSIONS

In this paper, we presented a novel model for Audio Chord Estimation based on the conformer architecture, enhanced with a consonance-informed label smoothing strategy and a decomposition-based decoding scheme. The motivation for incorporating perceptual smoothing emerged from our inter-annotator agreement analysis, which employed non-binary distance metrics and revealed that annotation discrepancies often involve harmonically related chords. Building on these insights, we introduced a learning strategy that integrates consonance-weighted targets into the training process.

Experimental results show that the proposed model achieves strong performance across both standard and non-binary evaluation metrics, with notable gains in capturing fine-grained harmonic relationships. Additionally, the proposed decomposition decoder not only enables chord prediction without relying on a fixed chord vocabulary, but also contributes to consistent performance improvements.

[2] `https://github.com/andreamust/consonance-ACE`

6. ACKNOWLEDGMENTS

This work is supported by IA y Música: Cátedra en Inteligencia Artificial y Música (TSI-100929-2023-1), funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial, and the European Union-Next Generation EU, under the program Cátedras ENIA 2022 para la creación de cátedras universidad-empresa en IA, and IMPA: Multimodal AI for Audio Processing (PID2023-152250OB-I00), funded by the Ministry of Science, Innovation and Universities of the Spanish Government, the Agencia Estatal de Investigación (AEI) and co-financed by the European Union.

7. REFERENCES

[1] W. B. de Haas, F. Wiering, and R. C. Veltkamp, "A geometrical distance measure for determining the similarity of musical harmony," *Int. J. Multim. Inf. Retr.*, vol. 2, no. 3, pp. 189–202, 2013.

[2] J. de Berardinis, A. Meroño-Peñuela, A. Poltronieri, and V. Presutti, "The harmonic memory: a knowledge graph of harmonic patterns as a trustworthy framework for computational creativity," in *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, Y. Ding, J. Tang, J. F. Sequeda, L. Aroyo, C. Castillo, and G. Houben, Eds. ACM, 2023, pp. 3873–3882.

[3] Y. Huang, S. Lin, H. Wu, and Y. Li, "Music genre classification based on local feature selection using a self-adaptive harmony search algorithm," *Data Knowl. Eng.*, vol. 92, pp. 60–76, 2014.

[4] J. Pauwels, F. Kaiser, and G. Peeters, "Combining harmony-based and novelty-based approaches for structural segmentation," in *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013, Curitiba, Brazil, November 4-8, 2013*, A. de Souza Britto Jr., F. Gouyon, and S. Dixon, Eds., 2013, pp. 601–606.

[5] J. Pauwels, K. O'Hanlon, E. Gómez, and M. B. Sandler, "20 years of automatic chord recognition from audio," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 54–63.

[6] T. Carsault, J. Nika, and P. Esling, "Using musical relationships between chord labels in automatic chord extraction tasks," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 18–25.

[7] 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," *Scientific Data*, vol. 10, no. 1, p. 641, Sep 2023.

[8] H. V. Koops, W. B. De Haas, J. A. Burgoyne, J. Bransen, A. Kent-Muller, and A. Volk, "Annotator subjectivity in harmony annotations of popular music," *Journal of New Music Research*, vol. 48, no. 3, p. 232–252, may 2019.

[9] T. de Clercq and D. Temperley, "A corpus analysis of rock harmony," *Popular Music*, vol. 30, no. 1, p. 47–70, 2011.

[10] A. Mcleod, X. Suermondt, Y. Rammos, S. Herff, and M. A. Rohrmeier, "Three metrics for musical chord label evaluation," in *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, ser. FIRE '22. New York, NY, USA: Association for Computing Machinery, 2023, p. 47–53.

[11] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[12] B. McFee and J. P. Bello, "Structured training for large-vocabulary chord recognition," 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. 188–194.

[13] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 5036–5040.

[14] N. C. Tamer, Y. Özer, M. Müller, and X. Serra, "High-resolution violin transcription using weak labels," in *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, A. Sarti, F. Antonacci, M. Sandler, P. Bestagini, S. Dixon, B. Liang, G. Richard, and J. Pauwels, Eds., 2023, pp. 223–230.

[15] M. Won, Y.-N. Hung, and D. Le, "A foundation model for music informatics," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1226–1230.

[16] A. Poltronieri, V. Presutti, and M. Rocamora, "Chordsync: Conformer-Based Alignment of Chord Annotations to Music Audio," in *Sound and Music Computing Conference - SMC 2024*, 2024.

[17] T. Fujishima, “Realtime chord recognition of musical sound: a system using common lisp music,” in *Proceedings of the 1999 International Computer Music Conference, ICMC 1999, Beijing, China, October 22-27, 1999*. Michigan Publishing, 1999.

[18] M. McVicar, R. Santos-Rodríguez, Y. Ni, and T. D. 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.

[19] M. Mauch and S. Dixon, “Approximate note transcription for the improved identification of difficult chords,” in *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, J. S. Downie and R. C. Veltkamp, Eds., 2010, pp. 135–140.

[20] E. J. Humphrey, T. Cho, and J. P. Bello, “Learning a robust tonnetz-space transform for automatic chord recognition,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*. IEEE, 2012, pp. 453–456.

[21] F. Korzeniowski and G. Widmer, “A fully convolutional deep auditory model for musical chord recognition,” in *26th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2016, Vietri sul Mare, Salerno, Italy, September 13-16, 2016*, F. A. N. Palmieri, A. Uncini, K. I. Diamantaras, and J. Larsen, Eds. IEEE, 2016, pp. 1–6.

[22] S. Sigtia, N. Boulanger-Lewandowski, and S. Dixon, “Audio chord recognition with a hybrid recurrent neural network,” in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, M. Müller and F. Wiering, Eds., 2015, pp. 127–133.

[23] 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, November 4-8, 2019*, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 644–651.

[24] 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, November 4-8, 2019*, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 620–627.

[25] Y. Wu and W. Li, “Automatic audio chord recognition with midi-trained deep feature and blstm-crf sequence decoding model,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 355–366, 2019.

[26] J. Deng and Y. Kwok, “Large vocabulary automatic chord estimation with an even chance training scheme,” 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. 531–536.

[27] L. O. Rowe and G. Tzanetakis, “Curriculum learning for imbalanced classification in large vocabulary automatic chord recognition,” in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, and A. Srinivasamurthy, Eds., 2021, pp. 586–593.

[28] J. A. Burgoyne, J. Wild, and I. Fujinaga, “An expert ground truth set for audio chord recognition and music analysis,” in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, A. Klapuri and C. Leider, Eds. University of Miami, 2011, pp. 633–638.

[29] H. V. Koops, W. B. de Haas, J. Bransen, and A. Volk, “Chord label personalization through deep learning of integrated harmonic interval-based representations,” *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1706.09552

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2818–2826.

[31] M. Buisson, P. Alonso-Jiménez, and D. Bogdanov, “Ambiguity modelling with label distribution learning for music classification,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 611–615.

[32] C. Liu and J. JaJa, “Class-similarity based label smoothing for confidence calibration,” in *Artificial Neural Networks and Machine Learning – ICANN 2021*, I. Farkaš, P. Masulli, S. Otte, and S. Wermter, Eds. Cham: Springer International Publishing, 2021, pp. 190–201.

[33] J. Lienen and E. Hüllermeier, “From label smoothing to label relaxation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 8583–8591, May 2021.

[34] C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, “Self-supervised learning with random-projection quantizer for speech recognition,” in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka,

L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 3915–3924.

[35] Q. T. Duong, D. H. Nguyen, B. T. Ta, N. M. Le, and V. H. Do, "Improving self-supervised audio representation based on contrastive learning with conformer encoder," in *Proceedings of the 11th International Symposium on Information and Communication Technology*, ser. SoICT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 270–275.

[36] Y. Chae, J. Koo, S. Lee, and K. Lee, "Exploiting time-frequency conformers for music audio enhancement," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 2362–2370.

[37] M. W. Akram, S. Dettori, V. Colla, and G. C. Buttazzo, "Chordformer: A conformer-based architecture for large-vocabulary audio chord recognition," 2025. [Online]. Available: https://arxiv.org/abs/2502.11840

[38] C. Harte, "Towards automatic extraction of harmony information from music signals," PhD thesis, Department of Electronic Engineering, Queen Mary University of London, 2010.

[39] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "Mir_eval: A transparent implementation of common mir metrics." in *Proceedings of the 15th International Conference on Music Information Retrieval*, 2014, pp. 367–372.

[40] K. Giannos and E. Cambouropoulos, "Symbolic encoding of simultaneities: Re-designing the general chord type representation," in *DLfM '21: 8th International Conference on Digital Libraries for Musicology, Virtual Conference, July 28-30, 2021*, C. Arthur, Ed. ACM, 2021, pp. 67–74.

[41] H. V. Koops, B. de Haas, J. A. Burgoyne, J. Bransen, A. Kent-Muller, and A. Volk, "Annotator subjectivity in harmony annotations of popular music," *Journal of New Music Research*, vol. 48, no. 3, pp. 232–252, 2019.

[42] C. Harte, M. B. Sandler, S. A. Abdallah, and E. Gómez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *ISMIR 2005, 6th International Conference on Music Information Retrieval, London, UK, 11-15 September 2005, Proceedings*, 2005, pp. 66–71.

[43] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler, "Omras2 metadata project 2009," in *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center, Kobe, Japan, October 26-30, 2009*, 2009.

[44] J. A. Burgoyne, J. Wild, and I. Fujinaga, "An expert ground truth set for audio chord recognition and music analysis," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, A. Klapuri and C. Leider, Eds. University of Miami, 2011, pp. 633–638.

[45] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *ISMIR 2002, 3rd International Conference on Music Information Retrieval, Paris, France, October 13-17, 2002, Proceedings*, 2002, pp. 287–288.

[46] P. Micikevicius, S. Narang, J. Alben, G. F. Diamos, E. Elsen, D. García, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.