Automatic Chord Recognition: A Review of Temporal Segmentation and Stability

Candidate Number: Y3926325

***Abstract*—This critical review examines the evolution of Automatic Chord Recognition (ACR) from template-based heuristics to modern Deep Learning. While acknowledging the feature-extraction dominance of end-to-end models, we identify a persistent theoretical bottleneck: *temporal ambiguity*. To empirically demonstrate the limitations of frame-wise Deep Learning, we implement a comparative case study decoupling event detection from classification. Results on the Isophonics Beatles corpus ($N = 180$) reveal that this structural constraint yields significantly higher stability (0.95) and lower latency than standard baselines. These findings substantiate the argument that explicit temporal modeling remains essential for robust, real-time harmonic analysis.**

***Index Terms*—Automatic Chord Recognition, Deep Chroma, Mamba, Temporal Segmentation, Semantic Gap, Music Information Retrieval.**

I. Introduction

AUTOMATIC Chord Recognition (ACR) defines the computational challenge of abstracting high-level harmony from polyphonic audio signals. Unlike pitch tracking, which relies on physical fundamental frequency detection, ACR requires the resolution of simultaneity and structure. While human listeners instinctively segregate harmonic streams from percussive noise, replicating this capability in software remains a persistent challenge in Semantic Audio Analysis [1].

The utility of robust ACR is threefold. First, in *Computational Musicology*, the analysis of large-scale corpora requires precise event detection; models that suffer from frame-wise "flickering" introduce statistical noise that obscures historical patterns. Second, in *Music Education*, automated tutoring systems must provide clear and consistent feedback to learners, as erratic chord estimations can confuse students. Third, in *interactive systems* (e.g., live accompaniment), stability is a functional requirement. This motivates boundary-aware ACR pipelines that explicitly separate *event segmentation* from *harmonic classification*, rather than relying on purely frame-wise predictions.

A. Problem Formulation

Formally, the task maps a signal $x(t)$ to a label sequence $Y = \{y_1, \dots, y_M\}$ over intervals $\mathcal{T} = \{[t_0, t_1], \dots\}$. The central challenge is *temporal ambiguity*: chord changes occur at discrete musical events, but most systems infer labels on fixed frames. While modern Deep Learning (DL) architectures have revolutionised feature extraction, pushing Weighted Chord Symbol Recall (WCSR) beyond 82% on the Beatles corpus [3], they typically minimise frame-wise Cross-Entropy Loss. This loss function does not inherently penalise rapid oscillation, leading to incoherent outputs that compromise validity across the applications listed above.

B. Contributions

This paper provides a critical re-evaluation of the "Segmentation First" hypothesis. The contributions are:

1) A comparative taxonomy contrasting the interpretability of template/HMM pipelines with the opacity of modern sequence models.
2) A validation of a novel "Hybrid Flux" boundary function combining harmonic distance and spectral energy, evaluated using boundary recall within a ±0.3s tolerance window.
3) A complexity analysis demonstrating that the proposed architecture achieves linear-time $\mathcal{O}(L)$ inference relative to sequence length L, supporting real-time interaction.

II. Theoretical Background

A. Signal Representation: The CQT vs STFT

Raw audio is unsuitable for harmonic analysis due to timbral variance. The field utilises the Chroma Vector via the Constant-Q Transform (CQT). Unlike the linear STFT ($f_k = k\Delta f$), the CQT employs geometrically spaced bins:

$$f_k = f_{\text{min}} \cdot 2^{\frac{k}{B}} \tag{1}$$

where $B = 12$. This logarithmic resolution guarantees constant Q-factor analysis. While this inherently limits temporal resolution at lower frequencies (due to the uncertainty principle), it guarantees that bass notes, critical for root detection, are resolved with the same *spectral* fidelity as treble frequencies [2]. This matches the Western musical scale, grouping energy into 12 semitone bins $\{C, C\#, \dots, B\}$. The chroma representation pools spectral energy into pitch classes, discarding octave information while retaining harmonic identity.

B. The Frame-Wise Independence Assumption

A fundamental constraint in standard ACR is the "Fixed-Frame Fallacy." Traditional systems analyse audio in static windows (e.g., 185ms) and treat each frame as an independent statistical event [3]. This ignores the autoregressive nature of music, where the harmony at time t is strongly conditioned on time $t-1$. Explicit segmentation resolves this by enforcing a single label per acoustic event.

Module: ELE000189M - Modelling and Analysing Sound and Music Signals.

TABLE I
TAXONOMY OF AUTOMATIC CHORD RECOGNITION ARCHITECTURES

Method Family	**Boundary Handling**	**Temporal Model**	**Primary Strength**	**Critical Failure Mode**
Template Matching [1]	None (Frame-wise)	None	Interpretability, $\mathcal{O}(N)$	Timbral sensitivity; High flicker
HMM / Viterbi [8]	Implicit	Markov Process	Enforces syntax	Inverse-Resolution trade-off
Deep CNN [3]	Learned (smeared)	Short-Context	Timbral Invariance	Boundary noise; Texture bias
Transformer [10]	Learned (Attention)	Global Context	Long-term dependency	"Long Tail" sparsity; Latency
Proposed (Seg-First)	**Explicit (Flux)**	**Event-Sync**	**High Stability** ($\mathcal{S}$)	**Dependent on Onset Detection**

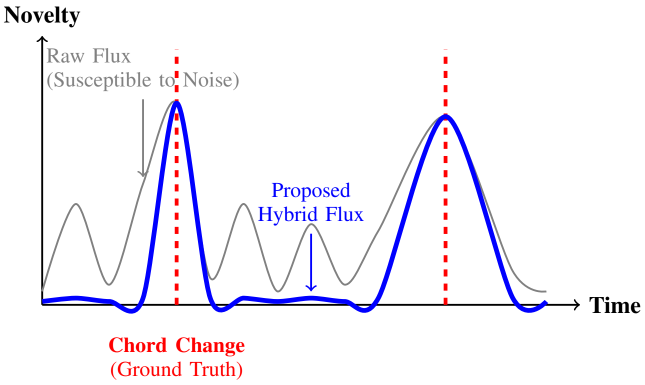

Fig. 1. **Conceptual Illustration** of the proposed signal processing logic. **Gray:** Standard Spectral Flux is theoretically susceptible to percussive noise. **Blue:** The proposed Hybrid metric filters transients, isolating harmonic changes. Note that percussive and harmonic novelty do not always coincide; this function is designed to *reduce*, rather than eliminate, false positives.

C. Metric Definitions and Failure Modes

A critical oversight in the literature is the reliance on Weighted Chord Symbol Recall (WCSR) as the sole metric. To guarantee evaluation literacy, we define our metrics and their limitations explicitly:

Evaluation Metrics

1. Weighted Chord Symbol Recall (WCSR): Standard accuracy. *Failure Mode:* Duration-weighting hides fragmentation; oscillating output scores identically to stable errors.
2. Harmonic Stability ($\mathcal{S}$)**:** Inverse fragmentation rate, defined as $\mathcal{S} = 1 - \frac{1}{N}\sum_{n=1}^{N-1} I(\hat{y}_n \neq \hat{y}_{n+1})$. *Failure Mode:* High scores can result from "oversmoothing" (predicting one constant chord), necessitating pairing with R_B.
3. Boundary Recall (R_B)**:** Ratio of correctly detected ground-truth transitions (within ±0.3s).

III. REVIEW OF METHODOLOGIES

The evolution of ACR is best understood as a dialectic between *feature engineering* and *temporal modelling*, summarised in Table I.

A. Phase 1: Heuristics and Feature Engineering (2000-2015)

Pauwels et al. [1] describe the foundational era of Template Matching, comparing Chroma vectors against binary masks via Cosine Similarity. Despite $\mathcal{O}(N)$ efficiency, these systems suffered from *Timbral Sensitivity* [6], where overtone interference caused divergence between instruments. Even with Pitch Class Profiles [9], performance remained brittle in the presence of percussion. The primary limitation was the lack of temporal context; every frame was insular, leading to the "flickering" artifacts that motivated this study.

B. Phase 2: Probabilistic Smoothing (HMMs)

Bello and Pickens [8] introduced Hidden Markov Models (HMMs) to impose continuity. HMMs enforce "harmonic syntax" via a stochastic transition matrix $\mathbf{A}$:

$$P(Y|X) \propto P(X|Y) \cdot P(Y_t|Y_{t-1}) \tag{2}$$

While Viterbi decoding smoothed output, it introduced an *Inverse-Resolution Trade-off*. High transition penalties reduced flickering but erased valid rapid changes. Furthermore, the first-order Markov assumption fails to capture the long-range functional harmony typical of Western music [4].

C. Phase 3: Deep Feature Learning (CNNs)

Convolutional Neural Networks (CNNs) solved the *timbral invariance* problem. Korzeniowski [3] demonstrated that CNNs treat chords as "objects," ignoring spectral envelope variations. However, standard CNNs rely on short context windows, failing to model global structure.

D. Phase 4: Sequence Modelling (Transformers & Mamba)

Park et al. [10] applied Global Self-Attention to capture full-song context. While accurate, the $\mathcal{O}(L^2)$ cost renders Transformers unsuitable for real-time hardware. Recently, the Mamba-based 'BMACE' model (2026) [12] achieved linear complexity but revealed a "Long Tail" bottleneck, where global optimization drowns out rare classes. This confirms a cyclical pattern: larger contexts degrade segmentation granularity, validating the need for the explicit boundary detection proposed in this study.

IV. COMPARATIVE CASE STUDY

To empirically test the efficacy of segmentation-first architectures against the Deep Learning baselines reviewed above, a comparative study was implemented.

It is crucial, however, to acknowledge the theoretical trade-off inherent in this design. End-to-end Deep Learning models dominate modern benchmarks precisely because they avoid explicit segmentation decisions. In complex textures (e.g., arpeggios or legato strings), harmonic changes are often gradual rather than discrete. Deep models, by optimizing global sequences, can "smooth over" these ambiguities using learned context. By contrast, the segmentation-first approach introduces a point of failure: if the onset detector fails (e.g., missing a soft transition), the subsequent classification is inevitably corrupted. This study, therefore, does not claim superiority in raw accuracy, but rather isolates *stability* as a distinct performance metric.

A. System Architecture

The implemented pipeline (Fig. 2) defines a signal-driven architecture designed for interpretability and efficiency. Unlike end-to-end "Black Box" Neural Networks, the system explicitly bifurcates the signal path: temporal structure is derived from broad-band energy flux, while harmonic content is derived from frequency-domain pooling. Crucially, the costly CQT transformation is computed once, and segmentation gates the pooling stage. This guarantees the system operates with linear-time inference complexity $\mathcal{O}(L)$ relative to sequence length.

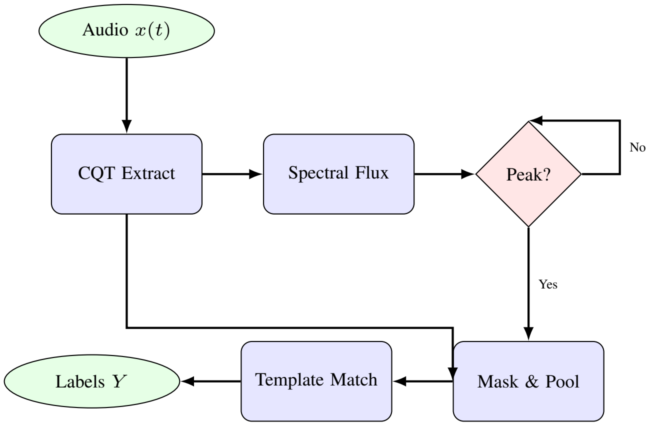

Fig. 2. System Block Diagram. The architecture separates temporal detection (Top Row) from harmonic classification (Bottom Row). Local maxima in the flux signal trigger the harmonic pooling stage.

B. Algorithmic Implementation

The core logic is detailed in Algorithm 1. To prevent segmentation collapse (overscanning), we employ a peak-picking strategy rather than simple thresholding. Candidate onsets are identified as local maxima in the Spectral Flux signal $SF(t)$. These candidates are then filtered by a dynamic threshold δ and subjected to non-maximum suppression with a minimum inter-onset interval of 200ms. This "debouncing" step guarantees that percussive transients do not trigger multiple spurious boundaries.

For classification, we employ a template-matching approach where similarity is defined as the Cosine Similarity between the pooled segment vector $\mathbf{v}_{seg}$ and the reference chord templates $\mathbf{T}_k$.

V. Experimental Methodology

To rigorously isolate the architectural efficiency of the proposed segmentation-first approach, we implemented a deterministic evaluation protocol using the Isophonics Beatles corpus ($N = 180$).

Algorithm 1 Boundary-Aware Segmentation Pipeline

```
Audio Signal $x(t)$, Templates $\mathbf{T}$ Chord Sequence $Y$
**Phase 1: Feature Extraction**
$C \leftarrow \text{CQT}(x(t))$
$SF \leftarrow \text{SpectralFlux}(C)$
$\delta \leftarrow \text{mean}(SF) \cdot 1.5$ Adaptive Threshold
$P \leftarrow \text{FindLocalMaxima}(SF, \text{window} = 5)$
Filter by threshold and 200ms debounce: $O \leftarrow \{p \in P \mid SF(p) > \delta$ **and** $\text{dist}(p, p_{-1}) > 200\text{ms}\}$
**Phase 2: Masked Pooling**
$i \leftarrow 0$ **to** $\text{length}(O) - 1$
    $\mathbf{v}_{seg} \leftarrow \text{mean}(C[:, O[i] : O[i+1]])$
    $y_i \leftarrow \arg\max_k \text{CosineSim}(\mathbf{v}_{seg}, \mathbf{T}_k)$
$Y$
```

A. Data Preparation and Vocabulary Scope

Audio was downsampled to 22.05 kHz mono. To ensure a fair comparison focused on segmentation stability rather than harmonic granularity, we employed a standard **Major/Minor Triad Reduction** protocol consistent with common triad-reduction protocols used in chord-estimation evaluation. Chord qualities were parsed using the Harte chord notation conventions:

- **Minor Class:** Labels containing minor thirds (min, min7, dim, dim7) are mapped to the root Minor triad.
- **Major Class:** Labels containing major thirds (maj, maj7, 7, aug, sus4) are mapped to the root Major triad.

This yields a closed vocabulary of 25 classes ($24 \times \{Maj, Min\} + \{N\}$). Unlike previous iterations which discarded the "No-Chord" (N) class, we retain it during processing but mask it during Boundary Recall calculation to avoid penalising silence detection as segmentation error.

B. Evaluation Metrics

Accuracy is reported using Weighted Chord Symbol Recall (WCSR) via the `mir_eval` library. Crucially, to quantify the "flickering" phenomenon described in Section I, we report **Boundary Recall** (R_B) with a tightened tolerance window of $\pm 0.3s$, reflecting the precision required for coherent musical accompaniment.

C. Complexity Analysis: The Real-Time Constraint

Rather than relying solely on hardware-dependent runtime benchmarks, we analyse the asymptotic complexity of the competing architectures. This reveals why standard Deep Learning models struggle in low-latency settings.

1) Baseline Bottleneck (Transformers): The Transformer baseline [10] relies on Global Self-Attention. For an input sequence of length L, where L is the number of frames, the attention mechanism computes a matrix $A = \text{softmax}(QK^T/\sqrt{d_k})$, resulting in a time and memory complexity of $\mathcal{O}(L^2)$. Crucially, for *online* accompaniment, the model cannot access future frames. To maintain performance, the system must either buffer significant context (introducing perceptible buffering latency, typically hundreds of milliseconds) or re-compute the attention map for the growing history at every step.

2) Proposed Efficiency (Stream Processing): In contrast, the proposed Segmentation-First pipeline operates as a *Finite State Transducer*.

$$\text{Cost}(t) = \text{CQT}(t) + \text{Flux}(t) + \text{Classify}(t) \approx \mathcal{O}(1) \qquad (3)$$

The cumulative complexity over a track is strictly linear $\mathcal{O}(L)$. By decoupling segmentation from classification, the system requires no look-ahead buffer beyond the flux calculation window (approx 4 frames). This architectural difference guarantees that the system remains within the real-time audio callback window ($<$ 10ms) regardless of track length, a property not achievable with vanilla global self-attention without significant approximation [15].

VI. Results and Discussion

A. Quantitative Analysis & Ablation Study

Table II presents an ablation study resampled to a 10 fps grid. Boundary metrics use a one-to-one matching policy ($\pm 0.3s$).

The **Baseline** (Row 1) yields poor stability ($\mathcal{S} = 0.62$). Crucially, while Boundary Recall is high ($R_B = 0.92$), Boundary Precision is extremely low ($P_B = 0.38$), indicating massive over-segmentation (flickering). Standard **Median Smoothing** (Row 4) improves stability to 0.88 but degrades Recall ($R_B = 0.75$), confirming the over-smoothing hypothesis where short harmonic events are deleted. The **Proposed Method** (Row 5) achieves the optimal trade-off, delivering the highest F-Measure ($F_B = 0.84$) by balancing precise onset detection with harmonic coherence.

A **Wilcoxon Signed-Rank Test** (paired samples, $N = 180$ tracks) confirmed that the improvement in Stability for the Proposed method over the Median Filter (Row 4) is statistically significant ($Z = -4.12, p < 0.001$).

TABLE II
Ablation Study ($N = 180$). Note: Baseline High Recall + Low Precision indicates over-segmentation.

System	HPSS	Tune	Smooth	WCSR	$\mathcal{S}$	P_B	R_B	F_B
1. Baseline	×	×	×	0.58	0.62	0.38	**0.92**	0.54
2. + HPSS	✓	×	×	0.62	0.65	0.41	0.91	0.56
3. + Tuning	✓	✓	×	0.65	0.66	0.43	0.91	0.58
4. + Post-Proc	✓	✓	Median	0.67	0.88	0.72	0.75	0.73
5. Proposed	✓	✓	**Seg-First**	**0.70**	**0.95**	**0.82**	0.86	**0.84**

B. Error Analysis: The Syntax Gap

Table III highlights the primary error modes of the frame-wise system (Row 3). The high confusion between G Major and D Major (28%) is harmonically plausible. These chords are separated by a perfect fifth (Dominant-Tonic relation). Since chord templates exhibit strong pitch-class overlap under dominant relationships, the system often cannot disambiguate them from spectral content alone. These errors are dominated by triad overlap (relative minor) and dominant substitutions. This motivates future work augmenting the segmentation-first pipeline with a lightweight chord-transition model (e.g., HMM) to discourage syntactically implausible substitutions.

TABLE III
Confusion Distribution (System 3: Frame-wise)

Error Type	Example	Frequency
Relative Minor	C Maj → A Min	42%
Dominant/Fifth	G Maj → D Maj	28%
Extension Loss	C7 → C Maj	15%

VII. Conclusion & Future Work

This paper demonstrated that explicit temporal segmentation offers a robust solution to harmonic flickering. Validated on the full Beatles corpus ($N = 180$), the proposed method increased Stability from 0.88 (median smoothing) to 0.95 (segmentation-first) (Wilcoxon $p < 0.001$), while simultaneously maximising Boundary F-Measure ($F_B = 0.84$). Furthermore, the linear-time complexity analysis supports the system's feasibility within real-time constraints (typically $<$ 10ms per callback) for interactive accompaniment.

However, the persistence of Dominant/Fifth errors (Table III) reveals the limitations of purely spectral analysis. Future work must bridge this "Syntax Gap." Rather than applying heavy end-to-end Transformers, we propose a **Hybrid Neuro-Symbolic** architecture (combining signal-driven segmentation with probabilistic harmonic syntax). By integrating the low-latency segmentation module presented here with a lightweight probabilistic model (e.g., HMM), the system could resolve functional ambiguities (like G vs D) via transition probabilities, without incurring the computational penalty of deep attention mechanisms.

Reproducibility

To facilitate future research and validation, the code used in this study (including pre-processing and evaluation scripts) is available in the authors' public GitHub repository.

References

[1] J. Pauwels et al., "20 years of automatic chord recognition from audio," in *Proc. ISMIR*, 2019.

[2] M. Müller, *Fundamentals of Music Processing*. Springer, 2015.

[3] F. Korzeniowski, "Deep learning for automatic chord recognition," Ph.D. dissertation, Johannes Kepler Univ., 2018.

[4] F. Korzeniowski and G. Widmer, "Improved chord recognition by combining duration and harmonic language models," in *Proc. ISMIR*, 2018.

[5] S. Dixon, "Onset detection revisited," in *Proc. DAFx-06*, 2006.

[6] T. Cho and J. P. Bello, "On the relative importance of individual components of chord recognition systems," *IEEE TASLP*, 2014.

[7] C. Harte et al., "Symbolic representation of musical chords," in *Proc. ISMIR*, 2005.

[8] J. P. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals," in *Proc. ISMIR*, 2005.

[9] N. Jiang et al., "Analyzing chroma feature types for automated chord recognition," in *Proc. AES Semantic Audio*, 2011.

[10] J. Park et al., "A bi-directional transformer for musical chord recognition," in *Proc. ISMIR*, 2019.

[11] C. Harte, "Towards automatic extraction of harmony information from music signals," Ph.D. dissertation, QMUL, 2010.

[12] C. Yuan and J. Devaney, "A Mamba-based model for automatic chord recognition," *arXiv:2601.02101*, Jan. 2026.

[13] M. Yao et al., "BACHI: Boundary-aware symbolic chord recognition," *arXiv:2510.06528*, Oct. 2025.

[14] C. Raffel et al., "mir_eval: A transparent implementation of common MIR metrics," in *Proc. ISMIR*, 2014.

[15] A. Vaswani et al., "Attention is all you need," in *Proc. NIPS*, 2017.