

ChordyV: A Hybrid Acoustic-Harmonic Audio Chord Estimation System with Metric Synchronization

Vie (Mulfie Ansyori Sofyan)

Independent Researcher / ChordyV Developer

masprodios.chordyv@gmail.com

ABSTRACT

This paper presents the submission of ChordyV to the MIREX (Music Information Retrieval Evaluation eXchange) Audio Chord Estimation task. Rather than relying on standard frame-level black-box neural networks that typically suffer from harmonic flickering and rhythmic misalignment, ChordyV introduces a multi-stage hybrid architecture. The system decomposes incoming audio into targeted acoustic components using deep stem separation, locks onto structural bar boundaries via deterministic physical-meter profiling, and extracts harmonic intervals using dual-layer Constant-Q Transform (CQT) arrays. Post-processing modules enforce music-theoretic constraints, passing-bass suppression, and global tonal center validation. The output strictly complies with Christopher Harte chord representation syntax and satisfies offline execution constraints required by the MIREX sandbox.

1. INTRODUCTION

Most conventional deep learning systems for Audio Chord Estimation (ACE) predict chord labels on short, fixed-duration temporal frames (e.g., 100 ms). This paradigm frequently produces boundary jitter, spurious transient chord assignments, and insensitivity to musical meter.

To overcome these shortcomings, ChordyV combines neural source separation with physics-based digital signal processing (DSP) and deterministic music theory. By aligning harmonic estimation to structural downbeats and metric grids before estimating pitch qualities, ChordyV ensures steady harmonic progressions that reflect authentic performance structures.

2. SYSTEM ARCHITECTURE

The ChordyV pipeline processes the audio stream through five sequential stages:

Raw Audio -> Stem Shaping -> Metric Locking -> Dual-CQT -> Harmonic Filtering -> Tonal Audit & Output

2.1. Neural Stem Ingestion & Spectral Shaping

The pipeline ingests raw audio and isolates fundamental musical roles using a multi-stem deep separation architecture:

- Root Subsystem: The extracted bass stem is combined with low-level polyphonic accompaniment and processed through a Butterworth low-pass filter to attenuate mid-to-high frequency vocal and melodic bleed, isolating clean fundamental bass contours.
- Harmonic Subsystem: Polyphonic instrumental stems (keyboards, guitars, and related instruments) undergo complementary high-pass filtering to remove low-frequency energy, preventing bass interference during upper triad identification.

2.2. Metric Induction and Phase Alignment (t0 Anchor)

Temporal profiling is conducted via multi-feature autocorrelation combining Harmonic Change Detection Functions (HCDF), low-frequency spectral flux, and transient onset envelopes:

- Meter Classification: The system evaluates harmonic and percussive lag ratios to distinguish between binary (e.g., 4/4) and compound or ternary (e.g., 6/8) time signatures dynamically.
- Downbeat Locking: The initial downbeat anchor (t_0) is resolved by optimizing transient coincidence across onset vectors, locking subsequent frame analysis to a structural bar grid.

2.3. Dual-Layer CQT Feature Extraction

Pitch tracking operates over two distinct acoustic registers to resolve bass movement and chord qualities independently:

- Fundamental Bass Register: A lower-octave logarithmic CQT array computes fundamental bass pitch-class distributions.
- Upper Harmonic Register: A dedicated mid-to-high CQT array captures upper chord textures, matched against prototypical templates covering standard diatonic and suspended triads.

2.4. Deterministic Filtering & Passing-Bass Elimination

Candidate chord assignments are refined through structural music-theoretic constraints:

- Passing-Note Suppression: Secondary bass excursions occurring on weak subdivisions are gated using spectral cosine similarity against the primary harmonic triad. Stepwise bass movements that maintain high chromatic coherence are classified as linear melodic bass lines rather than true harmonic changes, preventing over-segmentation.

- Inversion & Functional Mapping: Intervallic relationships between root and upper structures are mapped into functional pop representations, identifying natural inversions, suspensions, and chord extensions.

2.5. Tonal Auditor & Harte Formatting

Prior to generating the final annotations:

- Cadential Tonal Audit: The track-level sequence is evaluated against an algorithmic key-profile model that scores functional progressions (such as subdominant-to-dominant and dominant-to-tonic cadences) to stabilize isolated ambiguity.
- Harte Syntax Standardization: Chord symbols are formatted into standard Christopher Harte syntax (e.g., :maj, :min, :min7, :sus4, :dim, :hdim7). Slash chords are converted into relative scale-degree notation (e.g., G#:maj/2, C:maj/3) to guarantee strict compatibility with standard mir_eval evaluators.
- Temporal Continuity: Leading unvoiced segments (0.0 -> t0) and trailing boundaries are padded with non-harmonic tokens (N), ensuring an unbroken timeline across the entire file duration.

3. RUNTIME AND COMPLIANCE SPECIFICATIONS

- Offline Execution: Pre-trained model weights are pre-cached during the preparation phase (TORCH_HOME=/app/temp/chordyv_cache), guaranteeing deterministic inference inside isolated, network-restricted execution environments.
- Storage Safety: Intermediate waveform stems and processing buffers are automatically purged upon lab generation to prevent disk saturation during high-volume batch evaluations.

4. CONCLUSION

ChordyV bridges the gap between deep audio representation and deterministic musical grammar. By integrating metric synchronization, dual-register spectral isolation, and functional harmonic auditing, the framework provides an interpretable and structurally coherent solution for the Audio Chord Estimation task.

REFERENCES

1. C. Harte, M. Sandler, S. Abdallah, and E. Gómez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in Proc. 6th Int. Conf. on Music Information Retrieval (ISMIR), London, UK, 2005.
2. A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music Source Separation in the Waveform Domain," IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 29, pp. 642–652, 2021.
3. 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 Proc. 15th Int. Conf. on Music Information Retrieval (ISMIR), Taipei, Taiwan, 2014.