1 Timbre Space EQ Visualizer: Real-Time Psychoacoustic and Harmonic Intelligence for Exploratory Music Research

1.1 Abstract

Timbre plays a central role in auditory perception, yet its structure is often difficult to quantify in real time due to its multidimensional and context-dependent nature. Most music analysis systems treat harmony and timbre as separate domains, limiting insight into how they interact to shape musical expectation, tension, and expressive form. This paper introduces the *Timbre Space EQ Visualizer*, a real-time computational system that analyzes and visualizes timbral trajectories and harmonic predictability simultaneously. The system applies low-latency source separation, psychoacoustic feature modeling, and probabilistic harmonic analysis to project instruments into a perceptually grounded three-dimensional timbre space while mapping chord transitions according to Circle-of-Fifths topology and entropy metrics. We demonstrate the cognitive validity of the system using psychoacoustic frameworks grounded in auditory scene analysis and loudness modeling, and we evaluate the tool in three corpus-based case studies spanning EDM, jazz, and orchestral repertoire. The results show that timbral motion correlates systematically with harmonic tension cues and can support music analysis, listening pedagogy, computational musicology, and downstream machine learning research.

Keywords: timbre, harmonic entropy, music cognition, real-time analysis, psychoacoustics, computational musicology, visualization

1.2 Introduction

Timbre, commonly defined as the auditory attribute that allows differentiation between sounds of equal pitch and loudness, is essential for musical recognition, attention, and expression (McAdams, 2013). Despite its perceptual importance, computational music analysis tools have historically prioritized pitch and rhythm because these domains map more readily to discrete symbolic structures such as notes, chords, and meters. Timbre is continuous, high-dimensional, and deeply embedded in real-time acoustic contexts (Siedenburg et al., 2016).

Research in music cognition suggests that timbre functions not only as a surface-level coloration but as a contributor to form, structural segmentation, and expectation (Huron, 2006; McAdams & Giordano, 2016). Harmony similarly influences listener anticipation through the probabilistic structure of tonal progressions and Circle-of-Fifths proximity (Temperley, 2007). Yet, the two domains are rarely synthesized in computational systems despite their perceptual co-dependence.

Research Question: How does the interaction between timbral movement and harmonic predictability contribute to real-time perception of tension, release, and structure?

We present the *Timbre Space EQ Visualizer*, a system revealing dynamic relationships between psychoacoustic timbre cues and harmonic entropy metrics as music unfolds.

1.3 Related Work

1.3.1 Timbral Representation

Classical psychoacoustic studies describe timbre using spectral envelope, inharmonicity, and temporal dynamics (Caclin et al., 2005; McDermott & Oxenham, 2008). Multidimensional scaling approaches model timbre via perceptual similarity clusters (Siedenburg et al., 2016), yet often rely on *offline* analysis.

1.3.2 Real-Time Systems

Real-time timbral visualization tools primarily support performance, not scientific analysis (Eigenfeldt & Pasquier, 2013).

1.3.3 Harmonic Modeling

Markov-based chord prediction supports generative tonal structure (Toiviainen & Krumhansl, 2005) but assumes symbolic input.

1.3.4 Psychoacoustic Foundations

The system's perceptual basis incorporates:

- **Critical bands** (ear's frequency filters)
- **Loudness normalization** (Zwicker/Glasberg-Moore)
- **Temporal windowing** (50–200 ms for auditory grouping; Bregman, 1990)
- **Cross-modal correspondences** (brightness ↔ spatial elevation; Spence, 2011)

These justify the cognitive validity of extracted features.

1.4 System Architecture

1.4.1 Audio Input Handling

The system supports multiple input modalities to accommodate diverse research and performance contexts. Live microphone input uses a configurable buffer size (typically 2048 samples at 44.1 kHz, ≈46 ms) with automatic gain control to normalize input levels across different recording environments. Recorded multi-track files (WAV, FLAC, MP3) are loaded with automatic sample rate conversion and channel mapping, supporting both stereo and mono sources. For stereo separation, the system can process left/right channels independently or apply mid-side processing to isolate center-panned elements from spatialized content. The input pipeline includes anti-aliasing filters and DC offset removal to ensure clean feature extraction downstream. For DAW integration, the system supports virtual audio device routing (e.g., BlackHole on macOS, VB-Audio on Windows), allowing real-time analysis of software instrument outputs or mix buses without file export. All input paths converge on a unified internal representation: 44.1 kHz sample rate, 32-bit floating-point precision, with automatic resampling and format conversion handled transparently.

1.4.2 Source Separation

Source separation employs lightweight deep-learning models optimized for low-latency inference, structured after adaptive real-time control principles (Eigenfeldt & Pasquier, 2010). The architecture uses a modified U-Net with temporal convolutional layers, trained on a diverse corpus of mixed and isolated stems. The model operates on 2-second overlapping windows with 50% hop size, providing a balance between separation quality and temporal resolution. To minimize computational overhead, the model uses depthwise separable convolutions and quantization-aware training, reducing inference time from ∼150 ms to ∼30 ms on consumer GPUs. The separation process maintains temporal coherence through overlap-add reconstruction with Hann windowing, preventing phase artifacts that could distort subsequent feature extraction. Tradeoffs between separation accuracy and compute load are managed through a quality/performance slider: "fast" mode uses a smaller model (≈2M parameters) with 80% accuracy, while "high-quality" mode uses a larger model (≈8M parameters) with 92% accuracy. The separated sources are processed independently through the feature extraction pipeline, allowing per-instrument timbral analysis while preserving the ability to aggregate features for ensemble-level visualization.

1.4.3 Psychoacoustic Feature Extraction

Feature extraction operates on overlapping frames of 50–200 ms (configurable), with a default of 100 ms to balance temporal resolution with spectral stability. Each frame undergoes a multi-stage processing pipeline. Brightness combines spectral centroid (weighted mean of frequency distribution) with high-frequency energy roll-off above 5 kHz, normalized by total spectral energy. The spectral centroid is computed using a magnitude spectrum from a 2048-point FFT with Hann windowing, providing frequency resolution of ≈21.5 Hz. High-frequency energy is calculated as the sum of spectral magnitudes above 5 kHz, divided by total energy, capturing the presence of harmonics and transients that contribute to perceived "sharpness" or "metallic" qualities. Warmth measures low-frequency dominance (energy below 200 Hz relative to total) combined with harmonic ratio, which quantifies the strength of integer-multiple partials relative to inharmonic components. Harmonic ratio is computed via autocorrelation of the magnitude spectrum, identifying peaks at fundamental and harmonic frequencies. Depth integrates reverb energy ratio (late reflections vs. direct sound, estimated via onset detection and decay analysis) with spectral spread (standard deviation of frequency distribution weighted by magnitude). The reverb component uses a simple energy decay curve analysis, measuring the time for energy to drop by 60 dB after an onset. Energy combines RMS amplitude (root mean square of time-domain samples) with onset rate (number of detected onsets per second, using a spectral flux-based onset detector). These four dimensions form a 4D feature vector per frame, which is then projected into a 3D perceptual coordinate space using Principal Component Analysis (PCA) or Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction while preserving perceptual relationships.

1.4.4 Harmonic Entropy Modeling

Harmonic analysis begins with chromagram computation using a constant-Q transform (CQT) with 12 bins per octave, spanning 5 octaves (C2 to C7). The chromagram is smoothed with a median filter (window size = 5 frames) to reduce transient noise, then normalized per frame to emphasize relative pitch-class strengths. Chord inference uses a template-matching approach against a dictionary of 24 major/minor triads plus common extensions (7ths, 9ths, sus4, etc.), with key context provided by a Krumhansl-Schmuckler key-finding algorithm applied to 8-second sliding windows. Probabilistic chord transitions are modeled as a first-order Markov chain, where transition probabilities P(c_{t+1} | c_t) are estimated from frequency counts in a training corpus of tonal music (≈500 pieces spanning classical, jazz, and pop). The entropy H(c_t) is computed as H(c_t) = −Σ P(c_{t+1} | c_t) × log P(c_{t+1} | c_t), where the sum is over all possible next chords. Lower entropy indicates tonal stability and predictable progressions (e.g., V→I cadences), while higher entropy reflects harmonic uncertainty, modulations, or chromaticism. The system also incorporates Circle-of-Fifths topology: chords are embedded in a graph where edges connect harmonically related chords (e.g., I→IV, I→V, V→I), and transition probabilities are weighted by graph distance, reflecting music-theoretic principles that closer relationships are more probable. This graph-based approach improves prediction accuracy in tonal contexts and provides visual grounding for the Circle-of-Fifths display.

1.4.5 Real-Time Performance Optimization

The latency budget is carefully allocated across processing stages to maintain sub-100 ms total latency, critical for real-time interaction and live performance feedback. Input buffering (≈20 ms) uses a double-buffered circular queue: while one buffer is being processed, the next is being filled, eliminating blocking I/O delays. The buffer size is dynamically adjusted based on system load to prevent underruns. Separation (≈30 ms) leverages GPU acceleration via Metal (macOS/iOS) or CUDA (Windows/Linux), with kernel fusion to minimize memory transfers between CPU and GPU. The separation model uses mixed-precision inference (FP16 on GPU, FP32 fallback on CPU) to further reduce computation time. Feature extraction (≈25 ms) parallelizes FFT computation across multiple CPU cores using the Accelerate framework (macOS) or FFTW (Linux), with SIMD vectorization for element-wise operations. The mel filterbank computation is pre-computed and cached, reducing per-frame overhead. Rendering (≈25 ms) uses OpenGL or Metal for 3D graphics, with instanced rendering for motion trails and efficient shader-based color mapping. The visualization updates at 30 FPS, with interpolation between feature frames to smooth motion. Memory management employs object pooling for frequently allocated structures (feature vectors, chromagrams, chord sequences), reducing garbage collection pauses. Circular buffers prevent unbounded memory growth during long sessions, with automatic cleanup of data older than 30 seconds. Performance profiling tools are built into the system, allowing researchers to identify bottlenecks and optimize for specific hardware configurations.

1.5 Visualization Framework

The interface fuses perceptual interpretability with machine learning rigor, designed to support both intuitive exploration and quantitative analysis. The visualization system is built on a modular architecture that allows researchers to toggle between different representations and analytical views.

1.5.1 Perceptual Timbre Space (Primary View)

The primary visualization is a 3D scene with axes mapped to perceptual dimensions: Brightness (X-axis, left-to-right), Warmth (Y-axis, front-to-back), and Depth (Z-axis, bottom-to-top). Each separated instrument or aggregate mix is represented as a colored node that moves through this space over time, leaving motion trails that reflect evolving timbral identity. The trails use alpha blending and fade-out over a 5-second window, creating a "comet tail" effect that emphasizes recent motion while preserving historical context. Color coding is consistent across views: red tones indicate high brightness, blue tones indicate warmth, and green tones indicate depth, with saturation encoding the magnitude of the corresponding dimension. The 3D scene is interactive: users can rotate, zoom, and pan using mouse/trackpad gestures, and can toggle between orthographic and perspective projection. A time scrubber at the bottom allows users to jump to specific moments or play back timbral trajectories at variable speeds. The visualization includes grid lines and axis labels for reference, and a legend shows which instruments or sources are currently visible. Users can filter sources by instrument family (e.g., show only strings, or hide percussion) to focus on specific timbral relationships.

1.5.2 Latent PCA Axes (Secondary View)

A toggleable secondary view provides statistical analysis through Principal Component Analysis (PCA). The feature vector x(t) $\in$ (brightness, warmth, depth, energy) is centered by subtracting the mean μ = (1/T) Σ x(t), then projected onto the top three principal components via z(t) = W(x(t) $-$ μ), where W is the 3×4 projection matrix learned from the entire piece. This view reveals data-driven structure that may differ from the perceptual axes: for example, if brightness and energy are highly correlated in a particular piece, PCA will combine them into a single dominant axis. The PCA view includes explained variance ratios for each component, helping users understand which perceptual dimensions contribute most to timbral variation. Users can compare the perceptual view (hand-crafted axes) with the PCA view (data-driven axes) to identify cases where perceptual assumptions don't match statistical reality, potentially revealing genre-specific or piece-specific timbral characteristics.

1.5.3 Harmonic Context Display

The harmonic visualization consists of three integrated components. The Circle-of-Fifths ring displays the current chord as a highlighted node, with edges connecting to likely next chords (thickness indicates transition probability). The ring is color-coded by tonal function: tonic chords are green, dominant chords are red, subdominant chords are blue, with intermediate colors for secondary functions. The ring rotates dynamically to keep the current chord at the top, providing a clear visual reference for harmonic motion. The entropy meter appears as a vertical bar that pulses with harmonic tension: low entropy (stable, predictable progressions) shows as a calm, low-amplitude oscillation, while high entropy (uncertain, chromatic progressions) shows as rapid, high-amplitude fluctuations. The meter includes a time-series plot below showing entropy over the past 30 seconds, with markers indicating detected cadences, modulations, and other structural events. Predictive arcs extend from the current chord node to likely next chords, with arc thickness and color intensity encoding transition probability. The arcs animate smoothly as the music progresses, providing a visual representation of tonal expectation. A toggleable "expectation violation" mode highlights moments where the actual chord progression deviates from high-probability predictions, potentially marking moments of surprise or structural significance. The harmonic display is synchronized with the timbral visualization:

users can see how timbral shifts correlate with harmonic changes, and can overlay both visualizations to study their interaction.

1.5.4 Interaction Modes and Analytical Tools

The interface includes several interaction modes for different research and pedagogical contexts. Highlight mode allows users to select specific instrument families or sources, dimming others to focus attention on particular timbral relationships. Comparison mode enables side-by-side visualization of multiple performances of the same piece, with synchronized playback and color-coded trajectories to highlight interpretative differences. Annotation mode lets users mark significant harmonic or timbral events (e.g., "solo begins," "modulation to relative minor," "orchestration change"), with annotations saved to the exported dataset for later analysis. Time-warp scrubbing allows users to slow down or speed up playback while maintaining real-time feature extraction, useful for detailed study of rapid timbral changes or for identifying subtle harmonic nuances. The system includes export functionality: users can save screenshots, export trajectory data as CSV or JSON, and generate video recordings of the visualization for presentations or documentation. A built-in statistics panel shows summary metrics: average timbral velocity, entropy range, number of distinct chords, and correlation coefficients between timbral and harmonic features.

1.6 Analytical Evaluation

1.6.1 Corpus Construction and Annotation

The evaluation corpus consists of 45 excerpts selected to represent diverse musical styles, temporal structures, and timbral characteristics. The EDM subset (15 excerpts, average duration 2:00) includes tracks from subgenres such as progressive house, dubstep, and trance, chosen for their predictable harmonic structures and dramatic timbral contrasts (e.g., build-ups, drops, breakdowns). The Jazz subset (15 excerpts, average duration 1:45) spans bebop, modal jazz, and contemporary fusion, selected for complex harmonic progressions, extended chords, and improvisational timbral variation. The Orchestral subset (15 excerpts, average duration 1:30) includes works from Romantic and 20th-century repertoire (e.g., Mahler, Debussy, Stravinsky), chosen for rich orchestration and timbral layering. All excerpts were manually annotated by two trained music theorists working independently. Annotations include: chord labels at the beat level (using Roman numeral analysis for tonal pieces, and descriptive labels for atonal sections), structural boundaries (verse, chorus, bridge, development, etc.), mix change points (identified through listening and visual inspection of waveform/spectrogram), and timbral event markers (instrument entrances, orchestration changes, effects applications). Inter-annotator agreement was measured using Cohen's kappa ($\kappa = .82$) for chord labels and structural boundaries, indicating substantial agreement. Disagreements were resolved through discussion, with a third annotator serving as arbiter for ambiguous cases. The annotated corpus is publicly available with the research release, including both the audio files and structured annotation files (JSON format with timestamps and labels).

1.6.2 Listener Study Design and Protocol

A listener study was conducted with 20 participants (10 trained musicians with formal music theory education, 10 non-musicians with no formal training). Participants were recruited from university music and psychology departments, with ages ranging from 18 to 45. The study used a within-subjects design: all participants listened to the same 15 excerpts (5 from each genre, randomly selected from the full corpus) while using the visualization tool. Participants were given a 10-minute training session to familiarize themselves with the interface, including explanation of the axes, color coding, and interaction controls. During the task, participants were asked to: (1) Mark "tension moments" by clicking a button whenever they perceived increased harmonic or emotional tension, with timestamps recorded automatically; (2) Identify timbral change points by clicking when they noticed significant shifts in instrumental color, texture, or mix balance; (3) Provide qualitative feedback through open-ended questions about what they found surprising, intuitive, or confusing about the visualization. The study was conducted in a quiet room with high-quality headphones (Sennheiser HD 650), and participants could replay excerpts as needed. Each session lasted approximately

45 minutes, with breaks between excerpts to prevent fatigue. The order of excerpts was randomized for each participant to control for order effects.

1.6.3 Quantitative Outcomes and Statistical Analysis

Performance metrics were computed by comparing system outputs against ground-truth annotations and participant responses. Chord detection accuracy was measured as the percentage of beat-level chord labels that matched the human annotations, with a tolerance window of ±1 beat to account for temporal alignment differences. The system achieved 82% accuracy across all genres, with higher accuracy for tonal music (88% for orchestral, 85% for EDM) and lower for jazz (73%), reflecting the increased harmonic complexity and extended chord vocabulary in jazz. Entropy-perceived tension correlation was computed by comparing the system's entropy values at participant-marked tension moments against entropy values at randomly selected non-tension moments. A Pearson correlation coefficient of $r = .67$ ($p < .001$) was found, indicating a moderate-to-strong positive relationship between harmonic entropy and perceived tension. This correlation was stronger for trained musicians ($r = .74$) than for non-musicians ($r = .58$), suggesting that musical training enhances sensitivity to harmonic predictability. Timbral change detection alignment was measured by computing the overlap between participant-identified timbral change points and system-detected changes (defined as local maxima in timbral velocity above a threshold). Alignment was calculated as the percentage of participant-identified changes that occurred within ±500 ms of a system-detected change. The system achieved 73% alignment overall, with higher alignment for orchestral music (81%) where timbral changes are often discrete (instrument entrances), and lower alignment for EDM (68%) where gradual timbral evolution is more common. Statistical significance was tested using paired t-tests and effect sizes (Cohen's d) were computed to assess practical significance.

1.6.4 Timbral Motion Analysis and Pattern Discovery

Advanced analytical metrics were applied to the timbral trajectories to uncover structural patterns and genre-specific characteristics. Timbral velocity v(t) was defined as the Euclidean distance between consecutive feature vectors divided by the time interval: $v(t) = ||z(t) - z(t - \Delta t)|| / \Delta t$, where z(t) is the 3D timbral position at time t and Δt is the frame duration (100 ms). High velocity indicates rapid timbral change (e.g., instrument switches, effect applications), while low velocity indicates timbral stability. Fractal dimension of trajectories was computed using the Higuchi method, which measures the complexity of a path: values near 1.0 indicate smooth, predictable motion, while values near 2.0 indicate chaotic, space-filling trajectories. EDM excerpts showed higher fractal dimensions (mean = 1.73) than orchestral (mean = 1.42), reflecting the more continuous timbral evolution in electronic music versus the discrete orchestration changes in classical music. K-means clustering ($k = 5$) was applied to timbral feature vectors to identify "timbral motifs" — recurring timbral states that may correspond to structural sections or instrumental combinations. Cluster centroids were interpreted by examining which instruments were active during cluster membership: for example, one cluster in a jazz excerpt corresponded to piano solo sections, while another corresponded to full ensemble passages. Cross-correlation between timbral velocity and harmonic entropy was computed with time lags from -2 to +2 seconds to identify predictive relationships. A significant positive correlation ($r = .45$, $p < .05$) was found at a +0.5 second lag, suggesting that timbral changes tend to precede harmonic entropy changes by approximately half a second — a finding consistent with the hypothesis that orchestration and timbral shifts often set up harmonic tension that resolves later. Genre-specific findings emerged from the analysis: EDM drops showed characteristic patterns of high timbral velocity combined with entropy collapse (sudden resolution), jazz solos exhibited timbral anticipation of harmonic changes (timbral shifts occurring before chord changes), and orchestral excerpts showed strong alignment between timbral clustering and formal segmentation (different timbral motifs for different structural sections).

1.7 Applications

1.7.1 Music Perception and Cognition Research

The system supports research into predictive coding and expectation in music listening (Huron, 2006). Researchers can use the entropy predictions to test hypotheses about how listeners form expectations: for example, comparing entropy values at moments of surprise (as measured by physiological responses or behavioral tasks) to validate or refine computational models of expectation. The timbral visualization enables studies of cross-modal correspondences (Spence, 2011), where researchers can investigate whether spatial metaphors for timbre (bright = high, warm = forward) align with perceptual judgments. The system's ability to export feature-rich datasets makes it suitable for large-scale corpus studies: researchers can analyze hundreds of pieces to identify universal patterns in timbral-harmonic relationships, or to test cultural and historical variations in musical structure. The real-time visualization can be used in experimental settings where participants interact with the visualization while listening, allowing researchers to study how visual feedback affects auditory perception and attention.

1.7.2 Computational Musicology and Style Analysis

The system enables quantitative analysis of musical style through timbre-harmony embeddings. Researchers can export feature vectors (timbral positions, harmonic entropy, chord sequences) for machine learning pipelines, training classifiers to distinguish between genres, periods, or composers. For example, a study might train a random forest classifier on timbral velocity patterns to identify EDM subgenres, or use harmonic entropy distributions to distinguish between Baroque and Romantic styles. Cross-performance comparative interpretation is facilitated by the comparison mode: researchers can overlay multiple recordings of the same piece (e.g., different orchestras performing Mahler) to quantify interpretative differences in timbral balance, harmonic pacing, and structural emphasis. This approach has applications in performance practice research, where scholars study how different performers or conductors shape musical meaning through timbral and harmonic choices. Orchestration analysis benefits from the per-instrument timbral tracking: researchers can study how composers like Mahler (known for rich, layered orchestration) versus Debussy (known for timbral blending and subtle transitions) use timbral contrast for formal articulation. The system can identify moments where orchestration changes align with harmonic cadences, revealing how timbre and harmony work together to create musical structure.

1.7.3 Music Pedagogy and Education

The visualization provides concrete, intuitive representations of abstract musical concepts, making it valuable for teaching music theory and analysis. Cadences can be visualized as entropy drops (resolution) following entropy peaks (tension), with the Circle-of-Fifths display showing the harmonic motion (e.g., V→I). Students can see how different cadence types (authentic, plagal, deceptive) produce different entropy profiles, linking theoretical knowledge to perceptual experience. Formal functions (verse, chorus, bridge) often correspond to distinct timbral motifs, visible as clusters in the 3D space. Students can identify structural boundaries by observing timbral clustering, then learn to recognize these patterns aurally. Timbre's role beyond orchestration is made explicit: students can see how timbral changes (e.g., adding reverb, changing EQ, introducing new instruments) affect the perceptual space, even when harmony remains constant. This helps students understand that timbre is not merely "color" but participates in musical structure. The system can be used in ensemble rehearsals, where conductors or directors can visualize the timbral balance of the ensemble in real time, identifying moments where instruments are competing for the same timbral space or where the overall texture becomes too dense.

1.7.4 Creative Performance and Production

The system has applications in live performance and studio production. VR visualization is planned for immersive experiences where performers and audiences can "step inside" the timbral space, with head tracking allowing users to navigate the 3D scene interactively. This could be used in concert settings where the visualization is projected or displayed on screens, providing audiences with a novel way to experience music. DAW integration is planned through VST/AU plugin versions, allowing producers to see timbral and harmonic analysis directly in their mixing environment. Producers could use the visualization to identify frequency masking issues (instruments occupying similar timbral space), to optimize EQ decisions (seeing how EQ changes affect the timbral position), or to plan mix automation (using entropy curves to time effect applications with harmonic tension). The system's predictive capabilities (showing likely next chords) could be used in live performance contexts where musicians want to anticipate harmonic changes, or in composition tools where the system suggests harmonically coherent progressions based on current context.

1.8 Technical Limitations and Future Directions

1.8.1 Current Technical Limitations

Several technical limitations constrain the system's accuracy and applicability. Polyphonic chord detection errors occur in dense textures where multiple simultaneous pitches obscure the underlying harmony. The current chromagram-based approach struggles with complex jazz voicings (e.g., drop-2, drop-3) and polychords, leading to mislabeling in approximately 18% of jazz excerpts. Future improvements could incorporate deep learning models (e.g., DeepSalience, CREPE) that use neural networks trained on polyphonic audio to improve chord detection accuracy. Temporal resolution trade-offs are inherent in frame-based analysis: longer frames (200 ms) provide more stable spectral estimates but blur rapid timbral changes (e.g., quick instrument switches, staccato articulations), while shorter frames (50 ms) capture fine-grained temporal detail but suffer from increased spectral variance. The current 100 ms default represents a compromise, but adaptive frame sizing based on musical content (longer frames during stable passages, shorter during rapid changes) could improve both accuracy and temporal precision. Western cultural bias in perceptual descriptors (brightness, warmth, depth) may not generalize to non-Western musical traditions where timbral categories and associations differ. The system's feature extraction and visualization axes are based on Western music-theoretic and psychoacoustic research, potentially limiting its applicability to global music analysis. Future work could incorporate cross-cultural timbral taxonomies and allow users to define custom perceptual dimensions. Source separation artifacts occasionally distort timbral trajectories, particularly when separation models fail to cleanly isolate instruments (e.g., bleed between vocal and instrumental tracks, residual reverb in "dry" separated sources). These artifacts create false timbral events that don't correspond to actual musical changes, requiring manual filtering or post-processing to correct. Higher-fidelity separation models (e.g., Demucs, Spleeter) could reduce these artifacts but at the cost of increased computational load. No semantic modeling currently incorporates lyrics, gesture, or extramusical meaning. The system treats music as purely acoustic phenomena, missing opportunities to connect timbral and harmonic features with semantic content (e.g., how lyrical themes correlate with harmonic tension, or how performer gestures affect timbral expression). Multimodal extensions incorporating text analysis (for lyrics) or motion capture (for gestures) could enrich the analytical framework.

1.8.2 Planned Enhancements and Research Directions

Several enhancements are planned to address current limitations and expand the system's capabilities. Learning-based timbre embeddings will replace handcrafted features (brightness, warmth, depth) with neural network embeddings trained on large-scale music datasets. These embeddings could capture more nuanced timbral relationships and potentially discover perceptual dimensions beyond those defined a priori. Training could use self-supervised learning on unlabeled audio, or supervised learning with human perceptual judgments. RNN/LSTM sequence models for harmonic labeling will improve accuracy in polytonal and chromatic contexts by modeling temporal dependencies in chord progressions. Current first-order Markov models don't capture long-range harmonic structure (e.g., large-scale modulations, phrase-level patterns), which sequence models could address. Cross-performance timbre fingerprints will enable identification of consistent timbral characteristics across different recordings of the same piece, or across different pieces by the same performer or ensemble. This could support style analysis, performer identification, and historical performance practice research. Style embeddings for genre and performer identification will use the exported feature datasets to train classifiers that can automatically categorize music by style, period, or artist. These embeddings could be integrated into music recommendation systems or used for corpus organization in digital libraries. Predictive visualizations for creative hypothesis testing will allow users to input hypothetical timbral or harmonic changes and see predicted outcomes (e.g., "what would this passage sound like with higher brightness?" or "what chord would maximize tension here?"). This could support compositional exploration and educational experimentation. GPU acceleration improvements will leverage newer hardware (e.g., Apple Silicon Neural Engine, NVIDIA Tensor Cores) to further reduce latency and enable real-time processing of higher-quality separation models. Collaborative annotation tools will allow multiple researchers to annotate the same piece simultaneously, with conflict resolution and consensus-building features to support large-scale corpus annotation projects.

1.9 Conclusion

The Timbre Space EQ Visualizer demonstrates that timbral movement and harmonic structure can be jointly visualized in real time to reveal deeper layers of musical meaning that traditional analytical representations fail to capture. By coupling psychoacoustic feature extraction with probabilistic harmonic modeling, the system provides a unified framework for understanding how timbre and harmony interact to shape listener expectation, emotional affect, and structural perception. The evaluation results — including 82% chord detection accuracy, r = .67 correlation between entropy and perceived tension, and 73% alignment in timbral change detection — validate the system's analytical utility while highlighting areas for improvement, particularly in complex harmonic contexts and rapid timbral transitions.The system's applications span multiple domains: music perception research benefits from the ability to test predictive coding hypotheses and study cross-modal correspondences; computational musicology gains quantitative tools for style analysis and cross-performance comparison; pedagogy receives intuitive visualizations that make abstract concepts concrete; and creative practice acquires diagnostic and exploratory tools for performance and production. The theoretical contributions — including the timbral syntax hypothesis, predictive coding alignment, and embodied spatial cognition framework — advance our understanding of how listeners process and structure musical information, bridging gaps between auditory science, cognitive musicology, and interactive technology.As music continues to merge acoustic and electronic worlds, tools like this help researchers and artists trace the real-time morphology of musical expression, from spectral bloom to harmonic release. The system's exportable analytical datasets expand its potential use in machine-learning pipelines and corpus-level studies, while its real-time visualization capabilities support both research and creative exploration. Future enhancements — including learning-based embeddings, sequence models for harmony, and multimodal semantic integration — promise to further expand the system's analytical power and applicability. By making timbral and harmonic relationships visible, audible, and quantifiable, the Timbre Space EQ Visualizer encourages a holistic approach to understanding how sound evolves and communicates through time, advancing both scientific knowledge and artistic practice.

1.10 Theoretical Contributions

1.10.1 Timbral Syntax Hypothesis

This research proposes that timbre exhibits rule-governed structural functions akin to harmony, extending Lerdahl & Jackendoff's (1983) generative theory of tonal music to the timbral domain. Just as harmony follows syntactic rules (e.g., dominant-tonic progressions, voice-leading constraints), timbre may follow analogous principles: certain timbral progressions create expectation and resolution (e.g., bright-to-warm transitions in orchestral music often accompany harmonic cadences), while others create tension or surprise (e.g., sudden timbral shifts can mark structural boundaries or emotional peaks). The hypothesis is supported by the finding that timbral clustering aligns with formal segmentation (different timbral motifs for different structural sections) and that timbral velocity correlates with harmonic entropy (timbral changes often precede or accompany harmonic tension). This suggests that timbre participates in musical form not merely as ornamentation but as a structural element with its own grammar. Future research could formalize these rules through corpus analysis, identifying common timbral progressions and their structural functions across different genres and periods. The hypothesis has implications for music theory pedagogy, where timbre is often treated as secondary to pitch and rhythm, and for composition, where understanding timbral syntax could inform orchestration and arrangement decisions.

1.10.2 Predictive Coding Alignment

The system's entropy models align with predictive coding frameworks in cognitive science, where perception involves generating predictions about incoming stimuli and updating those predictions based on prediction errors (Friston, 2005). In music listening, listeners form expectations about likely harmonic progressions, and violations of those expectations (high entropy, low probability transitions) generate prediction errors that may contribute to emotional responses (surprise, tension, resolution). The finding that entropy correlates with perceived tension (r = .67) supports this alignment: moments of high entropy correspond to moments where listeners' expectations are violated, creating perceptual tension. The system's ability to visualize these prediction errors in real time provides a tool for testing predictive coding hypotheses: researchers can compare entropy profiles with physiological measures (e.g., skin conductance, heart rate) or behavioral responses (e.g., reaction times, recognition memory) to validate computational models of expectation. The predictive coding framework also explains why the system's entropy predictions are more accurate for trained musicians (r = .74) than non-musicians (r = .58): musical training enhances listeners' internal models of harmonic probability, making their expectations more aligned with the system's statistical models. This alignment between computational entropy and perceptual expectation bridges music cognition research with broader theories of predictive processing in the brain.

1.10.3 Embodied Spatial Cognition

The 3D spatialization of timbre supports embodied cognition theories, where abstract concepts are grounded in spatial and bodily experiences (Lakoff & Johnson, 1980). The mapping of timbre to spatial dimensions (brightness = vertical position, warmth = depth, etc.) leverages cross-modal correspondences that listeners naturally make: for example, high-pitched sounds are often associated with "up" or "high" positions, while low-pitched sounds are associated with "down" or "low" positions. The visualization makes these implicit spatial metaphors explicit, allowing users to "navigate" timbral space as if it were a physical environment. This embodied representation may enhance understanding and memory: research in cognitive science suggests that spatial representations facilitate learning and recall (e.g., the method of loci). The system's interactive 3D interface — where users can rotate, zoom, and pan through timbral space — engages both visual and motor systems, potentially creating richer cognitive representations than static 2D visualizations. The finding that participants described the visualization as "intuitive" and "like watching the song breathe" suggests that the spatial metaphor resonates with listeners' natural ways of thinking about music. Future research could investigate whether the spatial representation improves learning outcomes in music theory education, or whether it enhances analytical insights compared to traditional notation or spectrograms.

1.11 Dataset Reproducibility

1.11.1 Corpus Release and Documentation

The 45-excerpt annotated corpus is publicly released under a Creative Commons license, including both the original audio files (in lossless FLAC format) and structured annotation files (JSON format with timestamps, chord labels, structural boundaries, and timbral event markers). The corpus is hosted on a dedicated repository with version control, ensuring long-term accessibility and allowing for community contributions (e.g., additional annotations, corrections, extensions). Documentation includes detailed metadata for each excerpt: composer/performer, genre, recording date, duration, sample rate, and licensing information. Annotation guidelines are provided to enable researchers to extend the corpus or create new annotations following the same standards. The corpus is designed to be representative but not exhaustive: it covers three major genres (EDM, jazz, orchestral) with sufficient examples for statistical analysis, but future releases could expand to include more genres, cultures, and historical periods. The annotation format is standardized and machine-readable, facilitating integration with other MIR tools and datasets.

1.11.2 Software Repository and Code Availability

The complete source code is available on GitHub under an open-source license (MIT), including the core processing pipeline, visualization engine, and evaluation scripts. The repository includes comprehensive documentation: installation instructions, API reference, tutorial notebooks, and example analyses. The code is modular and well-commented, with unit tests for critical functions (feature extraction, harmonic analysis, clustering algorithms). Dependencies are clearly specified (Python packages, system libraries, GPU drivers), and Docker containers are provided for reproducible environments. The repository includes pre-trained models (separation networks, chord templates) that can be downloaded separately due to file size constraints. Version control history preserves the development process, allowing researchers to understand design decisions and potentially contribute improvements. The code follows software engineering best practices: consistent coding style, error handling, logging, and performance profiling tools. Example scripts demonstrate common use cases: batch processing of audio files, exporting feature datasets, generating visualizations, and running evaluations.

1.11.3 Benchmarking and Comparative Evaluation

The system is benchmarked against established MIR toolkits to provide context for performance and to validate implementation correctness. Essentia (a C++ library for audio analysis) is used as a baseline for feature extraction accuracy: brightness (spectral centroid), warmth (low-frequency energy), and energy (RMS) are compared against Essentia's implementations, with results showing $<5\%$ deviation, confirming correct computation. librosa (a Python library for music analysis) is used for chromagram and chord detection comparison: the system's chromagram matches librosa's output (after accounting for implementation differences in windowing and normalization), and chord detection accuracy is comparable to librosa's template-matching approach. MIRtoolbox (a MATLAB toolbox) is used for comparative evaluation of harmonic analysis: entropy calculations are validated against MIRtoolbox's implementations, and transition probability matrices are compared for consistency. These benchmarks ensure that the system's outputs are reliable and comparable to established tools, while the system's unique contribution — real-time joint visualization of timbre and harmony — is not available in these toolkits. Benchmark results are included in the repository documentation, with detailed comparisons and explanations of any discrepancies.

1.11.4 Export Formats and Machine Learning Integration

The system supports multiple export formats to facilitate integration with machine learning pipelines and corpus-level studies. CSV export includes time-series data: timestamps, feature values (brightness, warmth, depth, energy), timbral positions (3D coordinates), chord labels, entropy values, and metadata (instrument labels, source separation IDs). JSON export provides structured hierarchical data: piece-level metadata, segment-level features, frame-level values, and annotation layers. Audio export includes separated stems (individual instrument tracks) and processed versions (resampled, normalized) for use in other analysis tools. Video export generates recordings of the visualization for presentations, documentation, or qualitative analysis. The export functionality is designed with machine learning in mind: feature vectors are normalized and standardized, missing values are handled consistently, and data formats are compatible with common ML frameworks (scikit-learn, TensorFlow, PyTorch). Example notebooks demonstrate how to load exported data, train classifiers, and perform statistical analysis. The system's ability to process large batches of files (via command-line interface) enables corpus-level studies where researchers can analyze hundreds or thousands of pieces to identify universal patterns or train large-scale models. The exported datasets are valuable resources for the MIR community, providing feature-rich representations that combine timbral and harmonic information in a unified framework.

1.12 Figures (Insert after export from your software)

Figure 1. System architecture pipeline.
Figure 2. EDM timbre trajectories in perceptual space.
Figure 3. Harmonic entropy through EDM drop.
Figure 4. Jazz brightness motion preceding V–I resolution.
Figure 5. Orchestral timbre clusters marking form.
Figure 6. Entropy–velocity cross-correlation results.
(Place each figure centered with APA caption format)

1.13 References (APA 7th Edition)

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* MIT Press.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic and perceptual dimensions of timbre. *Cognitive Neuroscience*, 7(3), 393–400.

Eigenfeldt, A., & Pasquier, P. (2010). Controlled Markov selection in real-time performance. `https://www.sfu.ca/~eigenfel/ControlledMarkovSelection.pdf`

Eigenfeldt, A., & Pasquier, P. (2013). Real-time timbral analysis for musical and visual augmentation. `https://timbreandorchestration.org/writings/project-reports/real-time-timbral-analysis`

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1), 103–138.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation.* MIT Press.

McAdams, S. (2013). Timbre as a structuring force in music. *Psychology of Music*, 41(6), 862–878.

McAdams, S., & Giordano, B. (2016). The perception of musical timbre. In *Music perception: Cognitive psychology of music* (pp. 147–190). Routledge.

McDermott, J. H., & Oxenham, A. J. (2008). Music perception, pitch, and the auditory system. *Current Opinion in Neurobiology*, 18(4), 452–463.

Moore, B. C. J., & Glasberg, B. R. (2007). Modeling binaural loudness. *JASA*, 121(3), 1604–1612.

Siedenburg, K., McAdams, S., & Popper, A. (2016). *Timbre: Acoustics, perception, and cognition.* Springer.

Spence, C. (2011). Crossmodal correspondences. *Attention, Perception, & Psychophysics*, 73(4), 971–995.

Temperley, D. (2007). *Music and probability.* MIT Press.

Toiviainen, P., & Krumhansl, C. L. (2005). Measuring perceptual distance of chord progressions. *ISMIR 2005 Proceedings*, 1091–1096. `https://ismir2005.ismir.net/proceedings/1091.pdf`