MixSense: AI Optimization for Contiguous Music Segmentation at Scale

Vipul Razdan
Email: vr7717@gmail.com

***Abstract*—This paper casts long-form music stream segmentation as an AI optimization problem over a self-similarity manifold, unifying evolutionary search for parameter discovery with globally optimal dynamic-programming inference to recover contiguous boundaries consistent with a track-count prior or a data-driven estimate. Starting from Fourier-derived spectral embeddings, the method constructs cosine self-similarity and time-aware cost surfaces that encode symmetry, contiguity, and evolutionary stability, then solves for the minimum-cost partition without heuristic change-point thresholds. The pipeline is learning-free yet intelligent, leveraging search and global reasoning instead of supervised labels, and is stress-tested on a hand-annotated corpus exceeding 640 hours with human-variance analysis to contextualize error and tolerance around true boundaries. Results show robust, scalable segmentation under both known and estimated segment counts, highlighting AI-style optimization as a powerful alternative to local novelty detectors and ad-hoc rules in music structure recovery.**

***Index Terms*—Music segmentation, dynamic programming, evolutionary algorithms, self-similarity matrices, DJ mixing, audio processing**

I. Introduction

Electronic Dance Music (EDM) represents a unique domain for audio segmentation research due to the prevalent practice of DJ mixing, where tracks are seamlessly blended to create continuous music streams. The fundamental challenge in segmenting such content lies in the intentional obfuscation of track boundaries through sophisticated mixing techniques that align beats, harmonize frequencies, and create smooth transitions [1]. Traditional segmentation approaches relying on local change-point detection often fail in this context because DJs deliberately minimize perceptual discontinuities.

The problem of contiguous segmentation differs significantly from standard clustering formulations. Where conventional clustering algorithms group similar items regardless of temporal ordering, contiguous segmentation requires that clusters appear sequentially in time (AAABBBCCC rather than ABCBACABC). This temporal constraint, combined with the recursive self-similar structure inherent in musical compositions, creates a complex optimization landscape that demands global reasoning rather than local heuristics [2]. Previous work in music structure analysis has predominantly focused on novelty detection through local similarity measures, but these approaches struggle when the number of segments is known a priori, as is often the case with published DJ tracklists.

Our research addresses this gap by formulating music segmentation as a constrained optimization problem where the objective is to find the optimal partition of a time series into a predetermined number of contiguous segments that maximize intra-segment similarity. This formulation leverages several key insights: musical tracks exhibit internal self-similarity through repeating patterns and structures; DJ mixing preserves harmonic and rhythmic continuity across transitions; and the global optimum can be efficiently computed using dynamic programming despite the combinatorial explosion of possible segmentations [3].

The contributions of this work include: (1) a comprehensive framework for contiguous music segmentation that operates effectively with or without prior knowledge of segment counts; (2) novel cost matrices that capture musical properties beyond simple similarity, including symmetry, contiguity, and evolutionary patterns; (3) an evolutionary optimization approach for parameter discovery that adapts to different mixing styles and musical genres; (4) extensive evaluation on a massive hand-annotated corpus of 640 hours of DJ mixes with analysis of human annotation variance; and (5) practical implementation considerations for real-world applications in music streaming and metadata generation.

The remainder of this paper is organized as follows: Section 2 reviews related work in music segmentation and structural analysis. Section 3 details our corpus and the challenges of human annotation in DJ-mixed content. Section 4 presents our feature extraction pipeline and self-similarity construction. Section 5 introduces our family of cost matrices and their musical motivations. Section 6 describes our dynamic programming solution and confidence estimation framework. Section 7 presents comprehensive experimental results, and Section 8 concludes with discussion of applications and future directions.

II. Related Work

Music structure analysis has evolved through several methodological paradigms, each with distinct strengths and limitations for segmentation tasks. Early approaches focused on detecting points of change through novelty functions derived from audio features. Foote [4] pioneered the use of self-similarity matrices with checkerboard kernels to identify structural boundaries, creating a one-dimensional novelty function that highlighted regions of significant change. While effective for detecting obvious transitions, these local methods lack global context and struggle when the number of segments is predetermined.

The use of self-similarity matrices for analyzing time-dependent data originated in dynamical systems with recur-

rence plots [5], but was quickly adopted for music information retrieval. Cooper and Foote [6] demonstrated how self-similarity matrices could reveal musical structure through visual patterns, inspiring numerous subsequent approaches that extract segmentation boundaries from these representations. However, most methods continued to rely on local peak detection in novelty functions rather than global optimization.

Hidden Markov Models (HMMs) and other stochastic approaches brought a probabilistic perspective to music segmentation. Levy and Sandler [2] used HMMs with Gaussian mixture models to represent segments as hidden states generating observed feature sequences. Plotz et al. [7] applied HMMs specifically to DJ-mixed music, though evaluation was limited to small corpora. While model-based approaches capture temporal dependencies well, they typically require training data and may overfit to specific musical styles or mixing patterns.

Dynamic programming emerged as a powerful technique for finding globally optimal segmentations under constraints. Goodwin and Laroche [3] formulated segmentation as a path-finding problem through a state graph modeling transition costs between segments. Their approach recognized that novelty peaks often occur within segments rather than at boundaries, necessitating global reasoning. However, their method didn't leverage the prior knowledge of segment count that's often available in DJ mixing scenarios.

Clustering-based methods adapted for temporal constraints offer another approach. Radu [8] developed a time-dependent modification of agglomerative clustering that only merges adjacent segments, while Badawy et al. [9] applied binary classification frameworks with threshold time horizons. These methods balance flexibility with constraints but may produce irregular segment sizes and lack the global optimality guarantees of dynamic programming.

Recent work has explored learning-based approaches with deep neural networks, but these require large annotated datasets and may not generalize well across the diverse mixing styles found in electronic music. Our method occupies a unique position by combining the global optimality of dynamic programming with adaptive parameter optimization through evolutionary search, creating a learning-free approach that leverages domain knowledge through carefully designed cost functions rather than supervised training.

III. Corpus and Human Annotation Analysis

The evaluation of music segmentation algorithms requires extensive, reliably annotated corpora that represent real-world listening scenarios. Our research utilizes a comprehensive collection of DJ-mixed radio shows totaling over 640 hours of audio content, substantially larger than corpora used in previous studies [9], [10]. The corpus includes three major trance music shows: "A State of Trance" by Armin van Buuren (198 hours), "Magic Island" by Roger Shah (198 hours), and "Trance Around the World" by Above & Beyond (162 hours), plus an additional 83 hours of mixes annotated by Mikael Lindgren.

Audio files were originally 44.1kHz, 16-bit stereo MP3s sampled at 192kbps, downsampled to 4000Hz 16-bit mono WAV files to reduce computational requirements while preserving perceptually relevant frequency information. The Nyquist theorem [11] dictates that this sampling rate captures frequencies up to 2000Hz, sufficient for analyzing the harmonic content of most musical instruments while discarding less informative high-frequency components.

A critical aspect of our research involves analyzing the reliability of human annotations, which serve as ground truth for evaluation. Through comparison of multiple independently created cue sheets for the same radio shows, we quantified human disagreement in boundary placement. Analysis of 115 shows with at least three distinct annotators revealed a standard deviation of 9.13 seconds in boundary placement, with distinctive peaks at 8-bar intervals (approximately 14.8 seconds at 135 BPM) corresponding to musical phrase boundaries.

This human variance establishes a practical upper bound on segmentation accuracy, as algorithms cannot reasonably outperform human consistency. As noted by domain expert Denis Goncharov, boundary placement in trance music involves subjective judgment about when a new track "becomes the focus of attention," particularly during gradual transitions where tracks overlap extensively. This ambiguity is an inherent property of DJ-mixed content rather than annotation error, necessitating tolerance margins in evaluation metrics.

The corpus exhibits approximately normal distribution of track lengths with a mean of 5.7 minutes, though shows vary in structure and mixing style. "Magic Island" features longer tracks on average (388.2 seconds) with more gradual transitions, while "A State of Trance" has shorter tracks (317.9 seconds) with potentially more complex mixing patterns. These differences highlight the need for adaptive algorithms that can accommodate varying musical styles and DJ preferences.

IV. Feature Extraction and Self-Similarity Construction

The foundation of our segmentation approach lies in transforming raw audio signals into meaningful feature representations that capture musical structure while remaining invariant to mixing artifacts. Our pipeline begins with preprocessing that downsamples audio to 4000Hz mono and divides it into contiguous, non-overlapping tiles of fixed duration. Tile size represents a trade-off between temporal resolution and frequency analysis quality, with larger tiles providing better frequency resolution but reduced time localization.

We employ Fourier analysis to transform time-domain audio tiles into frequency-domain representations using the discrete Fourier transform:

$$F(x_k) = X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi \frac{k}{N} n}$$

where $x_0, \ldots, x_N$ are complex input samples and $X_0, \ldots, X_N$ are their frequency-domain representations. The FFTW algorithm [12] efficiently computes this transformation, particularly when N is zero-padded to the next power of two.

We apply bandpass filtering to focus on perceptually relevant frequency ranges, typically between 50Hz and 1950Hz, capturing fundamental frequencies and lower harmonics of

TABLE I: Corpus descriptive statistics

Reference	DJ	Hours	Mean Tracks	Total Tracks	Shows	Avg. Length (s)
ASOT	Armin van Buuren	198	20.6	2247	109	317.9
MAGIC	Roger Shah	198	17.3	1839	106	388.2
TATW	Above & Beyond	162	20.1	1771	88	329.8
LINDMIK	Mikael Lindgren	83	25.0	900	36	331.1
Total		641	20.7	6757	339	341.7

most instruments while excluding less informative extreme frequencies. Let h and l represent high and low pass filters in Hz, with $\hat{h} = \lceil h \cdot \frac{N}{R} \rceil + 1$ and $\hat{l} = \lceil l \cdot \frac{N}{R} \rceil + 1$ representing their positions in the spectrum. The feature matrix D contains absolute values of the complex Fourier results within this frequency range.

To accentuate instrument harmonics and reduce sensitivity to absolute amplitude variations, we apply Gaussian first-derivative convolution filtering:

$$-\frac{2\tilde{\lambda}}{v^2} e^{-\frac{\tilde{\lambda}^2}{v^2}}$$

where $\tilde{\lambda} = \{-\lfloor 2v \rfloor, \lfloor -2v+1 \rfloor, \ldots, \lfloor 2v \rfloor\}$ and $v = b\frac{N}{R}$, with b representing filter bandwidth in Hz. This operation acts as an edge detector in the frequency domain, expanding harmonic peaks to ensure alignment across similar feature vectors.

The self-similarity matrix S is constructed from cosine distances between normalized feature vectors:

$$S_{ij} = 1 - \langle D_i, D_j \rangle$$

We apply normalizing transformations to enhance discriminative properties, first centering S around 0.5 by raising elements to the power $2s$, where $s = \frac{1}{T^2} \sum_{i,j=1}^{T} S_{ij}$, then applying a power transformation $\hat{c} \in [0.5, 1.5]$ followed by rescaling to $[-1, 1]$. These transformations balance positive and negative incentives for track placement while maintaining numerical stability.

V. Cost Matrices for Musical Segmentation

The core innovation of our approach lies in the design of specialized cost matrices that capture different aspects of musical structure beyond simple feature similarity. Each cost matrix evaluates candidate segments according to distinct musical principles, and their weighted combination enables robust segmentation across varying mixing styles and musical genres.

A. Summation Cost Matrix

The most straightforward cost matrix sums similarity values within candidate segments:

$$C(f, t, \omega, \bar{S}) = \sum_{i,j=f}^{t} \frac{\bar{S}}{(t - f + 1)^{\omega}}$$

where $\bar{S}$ incorporates an incentive bias parameter Ω that controls the balance between positive and negative values:

$$\bar{S} \leftarrow \hat{S}_{ij}(\Omega) = \begin{cases} \Omega S_{ij}, & \text{if } S_{ij} > 0 \\ (1 - \Omega) S_{ij}, & \text{otherwise} \end{cases}$$

Direct computation would require $O(TW^3)$ time, but we employ a dynamic programming recurrence that reduces this to $O(TW)$:

$$\tilde{C}(f, t) = \tilde{C}(f+1, t) + \tilde{C}(f, t-1) - \tilde{C}(f+1, t-1) + \hat{S}_{ft} + \hat{S}_{tf}$$

This formulation efficiently computes segment costs by leveraging overlapping subproblems, a key advantage of dynamic programming.

B. Symmetry Cost Matrix

Musical compositions, particularly in electronic dance music, often exhibit symmetric structures at multiple time scales. We capture this property through a symmetry cost matrix that evaluates mirror similarity along diagonals parallel to the minor diagonal of S. For diagonal $\Lambda(f, t, d)$ at distance d from the minor diagonal:

$$\Lambda(f, t, d) = \langle S_{f+d,f}, S_{f+d+1,f+1}, \ldots, S_{t,t-d} \rangle$$

The symmetry cost is computed as:

$$\bar{C}(f, t, \bar{\Omega})(\Lambda, \bar{\omega}) = \sum_{i=1}^{|\Lambda|} \frac{\delta(\Lambda_i, \Lambda_{|\Lambda|-i+1}, \bar{\Omega})}{i^{\bar{\omega}}}$$

where $\delta(p, q, \Omega)$ rewards symmetric pairs with the same sign and ignores those with different signs. This formulation captures the repetitive and symmetric nature of musical phrases while remaining computationally tractable through reuse of shorter interval costs.

C. Contiguity Cost Matrices

We develop two variants of contiguity cost matrices that capture different temporal patterns in musical structure. Static contiguity identifies horizontal traces in the similarity matrix indicating sustained self-similarity or self-dissimilarity, characteristic of repetitive musical sections. Evolutionary contiguity detects diagonal traces representing evolving musical patterns that change gradually over time.

The static contiguity algorithm applies nth-order differences in two dimensions to highlight contiguous regions, then weights diagonal elements by their distance from the main diagonal to emphasize longer-range dependencies. Evolutionary contiguity performs similar operations along diagonals parallel to the main diagonal, capturing progressive changes in musical content that maintain internal coherence while evolving over time.

Both contiguity matrices incorporate parameters for difference order, incentive bias, and normalization, allowing adaptation to different musical styles and mixing techniques.

The combination of these matrices with the summation and symmetry costs creates a comprehensive evaluation framework that addresses multiple aspects of musical structure simultaneously.

D. Gaussian Regularization

To incorporate prior knowledge about typical track lengths, we include a Gaussian cost matrix that favors segments the expected duration:

$$G(\varpi, N)_{tw} = e^{-\frac{1}{2}\frac{\varpi n}{\frac{1}{2}W}z} \quad \text{for } n = 1, 2, \ldots, W$$

This time-independent regularization prevents unrealistic segment lengths and improves robustness when other cost matrices provide ambiguous signals. The width parameter ϖ controls the tightness of the length constraint, with higher values enforcing stricter adherence to the expected duration.

VI. Dynamic Programming and Confidence Estimation

The segmentation problem with known track count m involves finding the optimal partition of T time points into m contiguous segments that minimizes the total cost. The number of possible segmentations grows combinatorially with T and m, making exhaustive search infeasible for realistic problem sizes. For a two-hour show with 25 tracks and minimum/maximum track lengths of 190 and 900 seconds, the number of valid segmentations exceeds 10^{56}.

We formulate the problem as finding an m/T-segmentation $\mathbf{t} = (t_1, \ldots, t_{m+1})$ satisfying $1 = t_1 < \ldots < t_m < t_{m+1} = T + 1$, where track i comprises times $\{t_i, \ldots, t_{i+1} - 1\}$. The loss of a segmentation is the sum of individual track costs:

$$\ell(\mathbf{t}) = \sum_{i=1}^{m} C(t_i, t_{i+1} - 1)$$

We compute the minimum loss $\mathcal{V}_m^T$ using dynamic programming with the recurrence:

$$\mathcal{V}_1^t = C(1, t)$$

$$\mathcal{V}_i^t = \min_{t_i} \left[C(t_i, t) + \mathcal{V}_{i-1}^{t_i - 1} \right] \quad \text{for } i \geq 2$$

where t_i ranges from $t - W$ to $t - w$, respecting track length constraints. This approach reduces time complexity from exponential to $O(TWm)$, making computation feasible for practical applications.

For applications requiring uncertainty quantification, we develop a confidence estimation framework based on posterior marginals of song boundaries. Fixing a learning rate η, we define the posterior probability that song j starts at time s as:

$$P(j, s) = \frac{\sum_{\mathbf{t} \in \mathbb{S}_m^T : t_j = s} e^{-\eta \ell(\mathbf{t})}}{\sum_{\mathbf{t} \in \mathbb{S}_m^T} e^{-\eta \ell(\mathbf{t})}}$$

This quantity can be efficiently computed using forward and backward recursions similar to the dynamic programming solution:

$$P(j, s) = \frac{\mathcal{H}_{j-1}^{s-1} \cdot \mathcal{T}_{m-j+1}^{s}}{\mathcal{H}_m^T}$$

$\mathcal{V}_0^4$	$\mathcal{V}_1^4$	$\mathcal{V}_2^4$	$\mathcal{V}_3^4$	$\mathcal{V}_4^4$
$\mathcal{V}_0^3$	$\mathcal{V}_1^3$	$\mathcal{V}_2^3$	$\mathcal{V}_3^3$	$\mathcal{V}_4^3$
$\mathcal{V}_0^2$	$\mathcal{V}_1^2$	$\mathcal{V}_2^2$	$\mathcal{V}_3^2$	$\mathcal{V}_4^2$
$\mathcal{V}_0^1$	$\mathcal{V}_1^1$	$\mathcal{V}_2^1$	$\mathcal{V}_3^1$	$\mathcal{V}_4^1$
$\mathcal{V}_0^0$	$\mathcal{V}_1^0$	$\mathcal{V}_2^0$	$\mathcal{V}_3^0$	$\mathcal{V}_4^0$

Dynamic Programming Table

Fig. 1: Dynamic programming approach to segmentation: the optimal partition is computed by combining solutions to overlapping subproblems, with track boundaries determined through backward reconstruction.

where $\mathcal{H}_m^t$ and $\mathcal{T}_m^f$ represent forward and backward partition functions computed through dynamic programming.

We derive two confidence measures from these posterior marginals. The track index confidence $\Psi(\bar{m})$ quantifies uncertainty in track ordering by comparing the two highest probability boundaries for each track:

$$\Psi(\bar{m}) = 1 - \frac{(\tilde{\zeta}_{\bar{m}})_2}{(\tilde{\zeta}_{\bar{m}})_1}$$

where $\zeta_i = P(j, \Pi(\mathcal{V}_m^T, j))$ and $\Pi(\mathcal{V}, i)$ returns the optimal boundary placement for track i. The time confidence $\xi(\bar{m})$ similarly measures uncertainty in boundary timing by comparing the two highest probability times for each boundary.

When the number of tracks is unknown a priori, we estimate it by finding the value Δ that minimizes the normalized cost:

$$n = \arg\min_{\Delta} \frac{\mathcal{V}_\Delta^T}{\Delta}$$

This approach identifies the "saddle point" where adding more segments no longer provides sufficient cost reduction, effectively balancing model complexity with fit quality.

VII. Experimental Evaluation

We conducted comprehensive experiments to evaluate segmentation performance under both known and estimated track counts, comparing our approach against established baselines including Foote's novelty detection method [4] and naive duration-based estimation.

A. Experimental Setup

We randomly selected six shows (two from each major radio program) as a training set for parameter optimization via evolutionary search. The genetic algorithm used a population size of 50, elite count of 7, crossover fraction of 0.5, and stopped when the objective function stalled for five generations. We optimized for both mean absolute error and median absolute error, testing five configurations of cost matrices: summation with Gaussian, symmetry with Gaussian, contiguity with Gaussian, evolution with Gaussian, and all matrices combined.

TABLE II: Segmentation results with known track count (mean-optimized parameters)

Cost Matrix	lindmik ⟨med, mean, std⟩	magic ⟨med, mean, std⟩	tatu ⟨med, mean, std⟩	asot ⟨med, mean, std⟩	All ⟨med, mean, std⟩
Contiguity	⟨14,70,133⟩	⟨9,19,53⟩	⟨11,35,94⟩	⟨13,59,121⟩	⟨12,43,103⟩
Evolution	⟨14,55,110⟩	⟨12,27,58⟩	⟨9,24,55⟩	⟨15,50,102⟩	⟨12,38,83⟩
Summation	⟨8,34,81⟩	⟨8,14,34⟩	⟨6,17,50⟩	⟨8,33,77⟩	⟨7,24,62⟩
Symmetry	⟨10,20,50⟩	⟨9,17,44⟩	⟨8,12,17⟩	⟨11,26,60⟩	⟨9,19,46⟩
All Combined	⟨8,20,54⟩	⟨8,17,51⟩	⟨6,9,14⟩	⟨7,24,58⟩	⟨7,18,48⟩

TABLE III: Track estimation and segmentation performance

Method	Correct Estimation Rate (%)	Precision (10s tolerance)	Recall (10s tolerance)	F1 Score (10s tolerance)
Our Method	45.7	0.65	0.61	0.63
Enhanced Foote	44.5	0.68	0.59	0.63
Naive Guess	11.5	0.52	0.48	0.50

Evaluation metrics included mean absolute error (robustness), median absolute error (typical accuracy), and standard deviation of errors (error spread). We also computed precision, recall, and F1 scores using tolerance thresholds from 5 to 30 seconds to account for human annotation variance.

B. Results with Known Track Count

Table II presents segmentation accuracy when the true number of tracks is provided to the algorithm. The combined cost matrix approach significantly outperforms individual matrices, reducing mean absolute error from 23.7 seconds (summation alone) to 17.6 seconds (all matrices). This improvement comes primarily from reduced catastrophic misplacements rather than improved timing precision, as the combined approach better maintains correct track ordering.

The "Trance Around The World" dataset achieved the best performance with median error of 6 seconds and mean error of 9 seconds, suggesting more consistent mixing patterns or clearer track boundaries. "A State of Trance" proved most challenging with mean error of 24 seconds, consistent with its higher human annotation variance. These differences highlight how mixing style and musical complexity affect segmentation difficulty.

Analysis of confidence measures revealed that uncertainty increases toward the middle of shows, where track ordering becomes more ambiguous. The combined cost matrix approach substantially reduced this mid-show uncertainty compared to individual matrices, explaining its improved robustness. Time placement confidence remained relatively stable throughout shows, indicating that boundary timing is generally less ambiguous than track ordering.

C. Results with Estimated Track Count

When the number of tracks must be estimated from the audio, our normalized cost approach correctly identified the track count in 45.7% of cases, compared to 44.5% for an enhanced Foote method (with radius constraint) and 11.5% for naive duration-based estimation. The enhanced Foote method rarely overestimated track count but frequently underestimated, while our approach showed more balanced errors.

For segmentation with estimated track counts, we evaluated performance using F1 scores with varying tolerance thresholds (Table III). Our method achieved 63% true positive rate at 10-second tolerance, comparable to Plotz et al.'s reported 81% [7] but on a substantially larger and more diverse corpus. Performance varied by dataset, with "Trance Around The World" achieving 72.2% true positive rate at 10 seconds while "A State of Trance" achieved only 58.1%.

The similarity between our method and the enhanced Foote approach for track estimation suggests that both capture fundamental aspects of musical structure, though through different mechanisms. However, when the track count is known, our method significantly outperforms Foote's approach, demonstrating the value of global optimization over local novelty detection for precise boundary placement.

VIII. Conclusion and Future Work

We have presented a comprehensive framework for contiguous segmentation of DJ-mixed music streams that combines evolutionary parameter optimization with globally optimal dynamic programming inference. Our approach effectively addresses the unique challenges of electronic dance music segmentation, where intentional boundary obfuscation through mixing techniques renders local change-point detection inadequate.

The key advantages of our method include: (1) robustness to varying mixing styles and musical genres through adaptive cost matrices; (2) computational efficiency enabling processing of two-hour shows in under two seconds; (3) flexibility to operate with or without prior knowledge of track counts; and (4) meaningful confidence estimates that identify ambiguous regions requiring manual review.

Practical applications of this technology include automated metadata generation for music streaming services, interactive track-seek interfaces for DJ mixes, and music structure analysis tools for producers and researchers. Services like SoundCloud and MixCloud could integrate our method to automatically segment uploaded mixes using published tracklists, enhancing user experience through precise track-level navigation.

Future work will focus on several directions: developing an online version that operates on streaming audio with limited

lookahead; implementing regularized hierarchical clustering for comparison with our dynamic programming approach; exploring deep feature representations that may capture musical structure more effectively than Fourier-based features; and extending the framework to multi-modal analysis incorporating lyrics, album art, and social metadata.

The fundamental insight of formulating music segmentation as constrained optimization over self-similarity manifolds has broader applicability beyond music information retrieval. Similar approaches could benefit video segmentation, document structure analysis, and any domain where contiguous segments exhibit internal consistency while evolving over time. By combining musical domain knowledge with efficient algorithms, our work demonstrates the power of AI optimization techniques for solving complex temporal segmentation problems at scale.

REFERENCES

[1] J. Foote, "Visualizing music and audio using self-similarity," *Proceedings of the seventh ACM international conference on Multimedia*, pp. 77–80, 1999.

[2] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, 2008, pp. 318–326.

[3] M. M. Goodwin and J. Laroche, "A dynamic programming approach to audio segmentation and speech/music discrimination," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2004, pp. iv–309.

[4] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *IEEE International Conference on Multimedia and Expo*, vol. 1, 2000, pp. 452–455.

[5] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *Europhysics Letters*, vol. 4, no. 9, p. 973, 1987.

[6] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis," in *ISMIR*, 2002.

[7] T. Plötz, G. A. Fink *et al.*, "Automatic detection of song changes in music mixes using stochastic models," in *18th International Conference on Pattern Recognition*, vol. 3, 2006, pp. 665–668.

[8] R. Curticapean, "Clustering-based audio segmentation with applications to music structure analysis."

[9] D. El Badawy, P. Marmaroli, and H. Lissek, "Audio novelty-based segmentation of music concerts," in *Proceedings of the 21st International Conference on Digital Audio Effects*, 2018.

[10] E. Peiszer, T. Lidy, and A. Rauber, "Automatic audio segmentation: Segment boundary and structure detection in popular music," in *Proc. of LSAS*, 2008.

[11] H. Nyquist, "Certain topics in telegraph transmission theory," *Transactions of the American Institute of Electrical Engineers*, vol. 47, no. 2, pp. 617–644, 1928.

[12] M. Frigo and S. G. Johnson, "The fftw web page," *Online: http://www.fftw.org*, 2004.