SoloSearch: A System for Content-Based Music Retrieval from Jazz Improvisation

Alif Ilham Madani
Cornell Tech
2 West Loop Road, New York
aim57@cornell.edu

***Abstract*—Identifying jazz songs from improvised solos is challenging due to constant melodic variation. We introduce "SoloSearch," a study comparing two retrieval methods: a statistical baseline using timbral features (MFCCs) and a deep learning Triplet CNN. Using a 3-option forced-choice protocol, we benchmarked these against human listeners. On the training domain (personal recordings), both algorithms achieved 100% accuracy, surpassing humans (94%). However, under domain shift (commercial recordings), the deep model collapsed to near-random performance (37.5%), while the simple baseline remained more robust (50%). Our findings show that on small datasets, deep networks suffer from acoustic overfitting—memorizing recording artifacts rather than musical structure—suggesting that simple feature engineering often generalizes better for data-scarce music retrieval.**

***Index Terms*—Music Retrieval, Jazz Improvisation, MFCC, Audio Fingerprinting, Machine Learning.**

I. Introduction

In the domain of jazz piano, no two performances of the same song are ever identical [1]. Solos, rhythms, and chord voicings constantly shift, creating a significant retrieval challenge when attempting to identify a specific song without remembering the title. This variability renders standard exact-match audio fingerprinting techniques, such as those used by Shazam, less effective as they rely on identical spectral peaks [2].

The problem addressed in this work is identifying the correct song title from a "Solo"—a highly improvised variation—based solely on its "Head" or main theme. This represents a distinct semantic gap: the system must match the *harmonic identity* of the song while ignoring substantial variations in tempo, key, and instrumentation.

In this paper, we present SoloSearch, a comparative study of retrieval systems designed to identify song titles from 10-second unlabeled audio clips. Our contributions are threefold:

1) A curated dataset of annotated jazz piano improvisations distinguishing between "Head" and "Solo" segments.
2) A rigorous benchmark comparing human performance (94% accuracy) against signal processing and deep learning approaches.
3) An analysis of "Domain Shift," demonstrating how deep learning models can fail to generalize on limited audio datasets compared to robust feature engineering.

The remainder of this paper details the data pipeline, the implementation of Non-learning vs. Learning-based algorithms, and a discussion on why simple statistical features outperformed deep neural networks on out-of-distribution tasks.

II. Methodology

To address the improvisational challenge, we developed a structured data pipeline consisting of seven distinct stages.

A. Data Ingestion and Annotation

The dataset comprises 8 raw personal jazz audio recordings, covering standards such as "Autumn Leaves," "Billie's Bounce," and "Misty." A critical step in the pipeline was precise manual annotation to distinguish between the "atlas" (the main theme or head of the song) and the "query" (the solo improvisation). Using timestamped annotations, we isolated the head and solo sections for each track.

B. Preprocessing

First, the solo improvisation sections were temporally split into training (70%) and testing (30%) segments before any chunking occurred. This "split-then-chunk" strategy was strictly enforced to prevent data leakage, ensuring that no overlapping audio frames could exist between the training and testing sets.

Subsequently, 10-second audio clips were generated using a sliding window approach with variable strides to optimize data density:

- **Reference (Atlas) & Training Queries:** A stride of 2.0 seconds was used (80% overlap) to maximize the number of available samples for the database and model training.
- **Test Queries:** A stride equal to the window size (10.0 seconds, 0% overlap) was used to simulate realistic, disjoint query scenarios and ensure fair evaluation.

Segments shorter than the 10-second window were automatically discarded to maintain input consistency.

C. Feature Extraction

We utilize 40 Mel-frequency cepstral coefficients (MFCCs) to sufficiently capture the rich timbral complexity and harmonic content of jazz piano music [3]. The MFCCs were computed directly from the raw audio waveform to preserve full timbral information, rather than isolating harmonic components. The resulting feature matrices, with a shape of $(40, T)$ where $T = 10$ seconds.

D. Algorithms

We implemented and compared two distinct algorithmic approaches to the jazz solo identification task.

1) Non-learning Method: This approach utilizes statistical feature aggregation to establish a performance baseline without the need for optimization or weight updates. For each audio clip, the time-varying MFCC matrix is averaged across the temporal axis, producing a single 40-dimensional feature vector $\mathbf{v}$ that summarizes the global timbral characteristics of the segment.

Classification is performed using a k-Nearest Neighbors (k-NN) approach ($k = 3$). The similarity between a query vector $\mathbf{v}_q$ and a reference vector $\mathbf{v}_r$ is calculated using cosine distance. This method assumes that despite tempo variations, the global timbral distribution of a song remains relatively distinct.

2) Learning-based Method: The learning-based approach employs a weight-sharing triplet network architecture [4]. Although the model is instantiated as a single convolutional encoder, during training it simultaneously processes three parallel input streams (Anchor, Positive, and Negative) through identical weights to optimize the embedding space via Triplet Margin Loss.

As described in Table I, the model consists of three 1D convolutional blocks. The network was trained using Triplet Margin Loss ($\alpha = 0.5$). For each training step, the data loader constructs a triplet (A, P, N) to minimize the Euclidean distance between matching pairs while maximizing the distance between non-matching pairs (Algorithm 1). Then, like the non-learning method, classification is performed using a 3-NN approach on the embedding space based on cosine distance.

TABLE I: 1D CNN Architecture for Harmonic Embedding

Layer Type	In Ch.	Out Ch.	Kernel	Output
Input (MFCC)	-	40	-	$(B, 40, T)$
Conv1D + ReLU	40	32	5	$(B, 32, T)$
MaxPool1D	32	32	2	$(B, 32, T/2)$
Conv1D + ReLU	32	64	5	$(B, 64, T/2)$
MaxPool1D	64	64	2	$(B, 64, T/4)$
Conv1D + ReLU	64	128	5	$(B, 128, T/4)$
Global AvgPool	128	128	-	$(B, 128)$
Linear (Dense)	128	64	-	$(B, 64)$
L2 Normalization	64	64	-	$(B, 64)$

To ensure generalization and robustness, we employed 5-fold cross-validation. A strict "safety gap" of 5 chunks (approx. 10 seconds) was dropped at the training-validation boundaries to eliminate temporal correlation leakage between the query clips used for training and those used for evaluation.

E. Evaluation

To strictly align the computational performance with the human benchmarking protocol, we implemented a unified evaluation framework for all methods (Non-learning, Learning-based, and Human).

We adopted the 3-Alternative Forced Choice (3-AFC) method. While the machine learning models are capable of performing open-set retrieval across the entire database, comparing this directly to human performance is infeasible due to the extreme cognitive load required for a human to mentally search a large musical catalog.

Consequently, for each test query, the search space is dynamically restricted to three candidates:

- One **Target** vector (the correct song from the reference atlas).
- Two **Distractor** vectors (randomly selected songs from the reference atlas).

This format reduces the task to a local relative comparison, allowing for a direct, apples-to-apples comparison between the biological and artificial systems. The model is considered correct if the calculated distance between the query and the Target is smaller than the distance to either Distractor. This establishes a clear probability-adjusted baseline ($Chance = 33.3\%$).

Algorithm 1: Triplet Training Procedure

Input: Query Dataset $\mathcal{Q}$ (Solos), Atlas Dataset $\mathcal{A}$ (Heads)
Output: Trained Embedding Model f_θ
Hyperparameters: Margin $\alpha = 0.5$, Learning Rate $\eta = 0.001$
Initialize model weights θ randomly
while *not converged* **do**
  Sample batch of Anchors $\{x_a\}$ from $\mathcal{Q}$
  foreach *anchor x_a with label y* **do**
    Sample Positive $x_p \sim \mathcal{A}$ s.t. $label(x_p) = y$
    Sample Negative $x_n \sim \mathcal{A}$ s.t. $label(x_n) \neq y$
  Compute embeddings:
  $E_a, E_p, E_n \leftarrow f_\theta(x_a), f_\theta(x_p), f_\theta(x_n)$
  Compute Triplet Loss:
  $\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} \max\left(0, \|E_a^{(i)} - E_p^{(i)}\|_2^2 - \|E_a^{(i)} - E_n^{(i)}\|_2^2 + \alpha\right)$
  Update weights:
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
return f_θ

III. Experimental Results

We evaluated the performance of both algorithmic approaches alongside the human benchmark. The experiments were conducted on two distinct datasets: the "Personal Recordings" (high acoustic similarity between query and reference) and "Popular Recordings" (high acoustic variance).

A. Human Performance Baseline

To establish a "Gold Standard" for difficulty, we conducted a 3-AFC perception challenge. Participants were presented with one anonymous jazz solo and asked to identify the correct song title from 3 reference "Head" options. This process was repeated for 8 distinct songs.

TABLE II: Human Perception Accuracy Benchmarks

Method	Accuracy
Random Chance	33.3%
Human (No Earphones)	56.0%
Human (With Earphones)	94.0%

As shown in Table II, acoustic clarity is the primary determinant of human success. Without isolation (earphones), human accuracy dropped to 56%, comparable to our computational baselines. Participants reported relying on secondary perceptible features—such as background noise floor, specific instrument timbres, and recording fidelity—rather than harmonic structure alone to make determinations.

B. Algorithmic Performance

The algorithmic evaluation followed the same 3-AFC protocol. We tested the models on two splits: the original *Personal* dataset (In-Distribution) and the external *Popular* dataset (Out-of-Distribution).

Both methods achieved perfect convergence when the training/reference split was included. The high performance here confirms that both the statistical aggregation (Non-learning) and the CNN (Learning) can effectively map queries to references when the acoustic channel characteristics are identical.

We visualized this relationship in Figure 1(a). As clearly depicted, the eight songs form distinct, isolated clusters. Crucially, the "Test Query" markers (represented by crosses) map directly onto the manifolds of their corresponding "Reference" clusters (represented by circles). For example, the test queries for *Misty* (blue crosses) are spatially indistinguishable from the reference tracks (blue circles). This overlapping distribution confirms that for personal recordings, the timbral features of the test solos are statistically identical to the reference heads.

To test true generalization, we evaluated the models on famous commercial recordings of the same songs. This introduced significant shifts in instrumentation, tempo, and production quality.

As illustrated in Table III, the accuracy dropped significantly. This failure is visualized in Figure 1(b). Unlike the personal recordings, the popular recording queries (crosses) exhibit significant "drift" away from their target centroids. In several instances, the query markers land in the sparse regions between clusters or seemingly associate with incorrect clusters (e.g., the *Yardbird Suite* query drifting towards *Misty*'s cluster). This visualizes the *domain shift* phenomenon: the algorithm is correctly measuring distance, but the "acoustic location" of the song has moved due to production differences.

The severity of the Learning-based method's failure is visualized in Figure 2. While the In-Distribution plot (Figure 2a) shows extremely tight, distinct clusters—evidence of strong supervision—the Out-of-Distribution plot (Figure 2b) reveals a collapse of this structure. Unlike the MFCC baseline, which maintained some proximity, the neural network projects the popular recording queries into completely incorrect regions of the embedding space, indicating that the model learned to recognize the *sound* of the recording environment rather than the *music* itself.

TABLE III: Algorithm Accuracy Comparison

Methodology	Personal (ID)	Popular (OOD)
Non-Learning Baseline	**100%**	**50%**
Learning-Based (CNN)	**100%**	37.5%

IV. Discussion

Our investigation into jazz solo retrieval yields two critical insights regarding the state of Music Information Retrieval (MIR) for improvisational content: the susceptibility of deep models to acoustic overfitting and the comparative robustness of simple statistical features under domain shift.

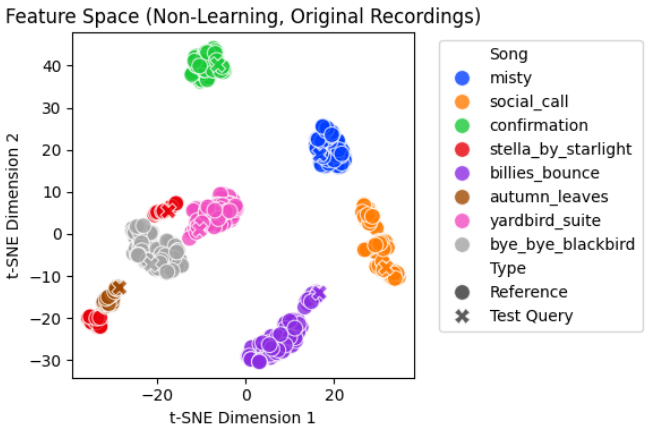
The most significant finding is the Learning-based method's performance collapse on the Out-of-Distribution (OOD) Popular dataset (37.5% accuracy), which barely exceeds random guessing (33.3%). This failure is visually confirmed by the "clustering collapse" in Figure 2. While the model produced tight, well-separated manifolds for the In-Distribution (Personal) recordings (Figure 2a), this structure completely disintegrated when applied to external recordings (Figure 2b).

This scattering indicates that the CNN did not learn to recognize the abstract harmonic identity of the songs. Instead, it solved the In-Distribution task by overfitting to "confounding" acoustic artifacts specific to the training domain—likely the room tone, microphone response, or silence profile. When presented with popular recordings lacking these specific artifacts, the model's learned weights penalized the correct matches, projecting them into irrelevant regions of the embedding space. This challenges the prevailing assumption that deep learning is invariably superior for audio classification; without massive-scale pre-training or data augmentation, these models identify a specific recording rather than the song structure.

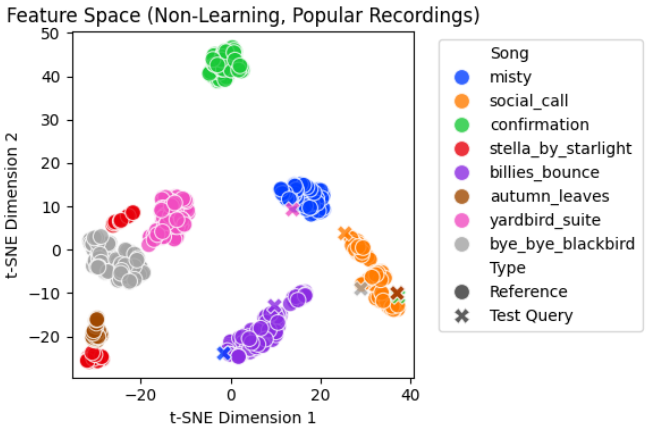
In contrast, the Non-learning method (Global MFCC Averaging) exhibited superior robustness on the OOD data (50% accuracy). By compressing the temporal dimension into a single statistical vector, this method inadvertently filtered out the high-frequency temporal noise that confused the CNN. While its absolute performance is not deployable, its ability to maintain some cluster proximity suggests that global timbral statistics are a more reliable proxy for song identity than complex, unregularized convolutional features when training data is scarce.

On the Personal dataset, human participants achieved **94% accuracy**, a score comparable to the Learning-based model's In-Distribution performance (**100%**). This demonstrates that for the specific task of identifying a musician's own recordings, the "Semantic Gap" is effectively closed; the machine is as capable as the human ear.

Conclusion: We have demonstrated that deep learning architectures can achieve super-human performance on In-Distribution jazz solo retrieval, effectively solving the problem of "session identification." However, this performance comes

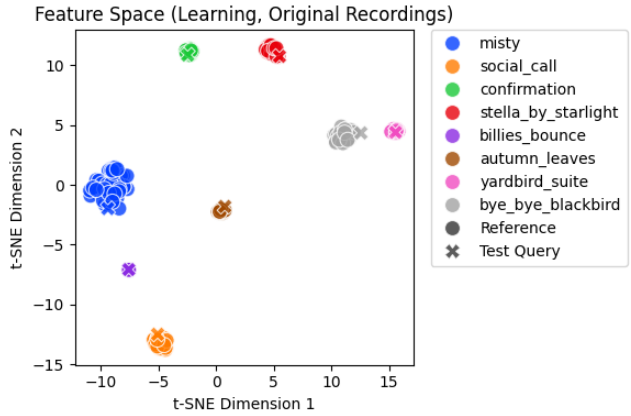


(a) **Non-Learning (In-Distribution):** Test queries (crosses) overlap perfectly with reference clusters (circles), indicating high acoustic similarity.

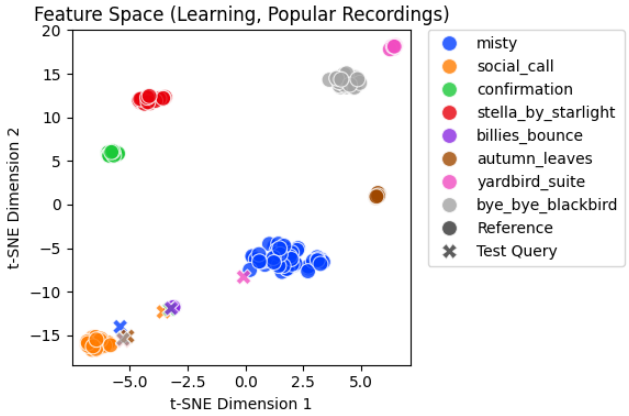


(b) **Non-Learning (Out-of-Distribution):** Test queries (crosses) drift away from their reference manifolds due to recording variations, causing classification errors.

Fig. 1: Comparative t-SNE visualization of the Non-learning feature space. Note the tight clustering in (a) versus the semantic drift in (b), which illustrates the ”Domain Shift” challenge discussed in Section V.



(a) **Learning (In-Distribution):** The Triplet network creates highly compact clusters for personal recordings. Test queries (crosses) are embedded almost identically to the references.



(b) **Learning (Out-of-Distribution):** The embedding space fractures on popular recordings. Note how the orange and pink test queries (crosses) are projected far from their target clusters, visually explaining the drop to 37.5% accuracy.

Fig. 2: Visualization of the Triplet Network's failure mode. While the model overfits perfectly to the training domain (a), it fails to learn invariant harmonic features, leading to severe scattering when applied to unseen production styles (b).

at the cost of extreme brittleness. For the broader task of generalized jazz standard identification, where the model must recognize a song across different artists and recording environments, our results suggest that current small-scale CNNs are insufficient. Future research must pivot toward invariant feature engineering (e.g., Chroma/Tonnetz) or large-scale pre-training to bridge the gap between identifying a *recording* and identifying a *song*.

References

[1] P. F. Berliner, *Thinking in Jazz: The Infinite Art of Improvisation*. University of Chicago Press, 1994.

[2] A. L.-C. Wang, “An industrial-strength audio search algorithm,” in *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR)*, 2003, pp. 7–13.

[3] B. Logan, “Mel frequency cepstral coefficients for music modeling,” in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, 2000.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.