Music Similarity Finder

Kanita Tafro
University of Sarajevo, Faculty of Electrical Engineering
Data Science and Artificial Intelligence
Sarajevo, Bosnia & Herzegovina
ktafro3@etf.unsa.ba

***Abstract*—Music similarity estimation is a central problem in Music Information Retrieval (MIR), with applications in retrieval, recommendation, and exploratory browsing. This paper presents a multi-layer music similarity framework that integrates psychoacoustic modeling, interpretable audio feature grouping, nonlinear dimensionality reduction, and melody-based querying. Low- and mid-level audio descriptors are first extracted from music recordings and perceptually biased using A-weighting to approximate human loudness sensitivity. The features are then organized into musically meaningful groups representing timbre, spectral characteristics, harmony, and rhythm, enabling explainable similarity analysis. To support visualization and exploratory interaction, Uniform Manifold Approximation and Projection (UMAP) is applied separately to each feature group, producing low-dimensional embeddings that preserve perceptual relationships in the data. Finally, a Query by Humming (QbH) component based on mel-spectral embeddings enables melody-driven retrieval that is robust to performance variability. The proposed architecture emphasizes interpretability, perceptual relevance, and modularity, providing a scalable foundation for music similarity exploration in academic and educational settings.**

I. INTRODUCTION

Music Information Retrieval (MIR) is an interdisciplinary research area encompassing computer science and information retrieval, musicology and music theory, audio engineering and digital signal processing, cognitive science, library science, publishing, and law. Its agenda, roughly, is to develop ways of managing collections of musical material for preservation, access, research, and other uses [1]. Peter Knees and Markus Schedl in their book "Music Similarity and Retrieval" [2] highlight three main paradigms of music information access: retrieval, browsing, and recommendation. *Retrieval* refers to scenarios in which a user formulates an explicit query—expressed through text, symbolic representations, or a piece of audio—to satisfy a specific information need, with the system returning relevant musical items or metadata, often in ranked form. *Browsing* describes an exploratory interaction mode in which the user has no precisely defined goal and incrementally navigates a music collection through iterative, user-driven interaction supported by intuitive interfaces. *Recommendation* is an automated process in which the system proactively suggests potentially relevant musical items based on explicit user preferences or implicitly inferred behavior, without requiring an active search query.

Across these three paradigms, music similarity serves as a core enabling concept, requiring systems to estimate how closely musical items relate in perceptual and structural terms. Because musical similarity is inherently subjective and multi-dimensional, MIR systems rely on audio-based representations that combine low-level acoustic descriptors with musically meaningful features capturing timbre, harmony, rhythm, and perceived loudness. Recent work emphasizes the integration of perceptually motivated feature weighting, interpretable feature organization, and embedding techniques to support both similarity computation and exploratory interaction. In this seminar paper, a multi-layer music similarity framework is presented that integrates psychoacoustic feature weighting, explainable feature grouping, nonlinear embedding, and Query by Humming, with the aim of providing an interpretable and modular approach to music similarity analysis grounded in established MIR principles.

II. RELATED WORK

The dataset used for this experiment is the GTZAN dataset, created by and named after George Tzanetakis [3], which contains 1000 half-minute music audio samples. The samples are categorized into 10 genres, with each genre containing 100 samples [4]. The original MARSYAS (**M**usic **A**nalysis and **R**etrieval **S**ystems for **A**udio **S**ignals) [5] server for the dataset download is known to be unstable and currently the only available source is a reupload of the dataset on Kaggle. The Kaggle reupload is the work of Andrada Olteanu, James Wiltshire, Lauren O'Hare and Minyu Lei[1] for a university project, as stated by the author Andrada in the dataset description. Alongside the reuploaded GTZAN dataset audio samples, the authors provided MEL spectrograms of each audio sample and two csv files of extracted numerical audio features. This experiment is inspired by the provided MEL spectrograms to use them as a baseline for the fourth layer. The given spectrograms aren't used but they are manually extracted using the Python library librosa in both npy and png format.

Aside from the server instability, Bob L. Sturm (2013) remarked distortions in audio samples [4] and the file *jazz.00054* was corrupted in the reuploaded Kaggle version. As this file can neither be played nor read by the Python library librosa, Kaggle users found the original distorted file and linked it in discussions under Andrada Olteanu's reupload. The corrupted file was replaced with the reuploaded file in discussions before working with the dataset.

[1]Dataset Kaggle download: "GTZAN Dataset - Music Genre Classification" kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification

Three original features (referred as layers in this paper) include A-weighting, feature grouping, and UMAPs. Query by humming was selected as the fourth bonus feature.

Layer 1 - A-weighted psychoacoustic features. Psychoacoustic modeling has long been used in audio analysis to bridge the gap between physical signal representations and human auditory perception. Fletcher and Munson's equal-loudness contours [6] established that perceived loudness varies nonlinearly across frequencies, forming the basis for standardized A-weighting curves used in audio measurement systems. In MIR, perceptually motivated representations have been shown to improve alignment with human similarity judgments by emphasizing mid-frequency regions critical to auditory sensitivity. Knees and Schedl (2016) [2] highlight the importance of perceptually grounded feature representations in similarity computation, while Peeters et al. (2011) demonstrate that perceptually informed spectral descriptors improve timbral modeling [7]. Applying A-weighting at the feature level, rather than the signal level, follows a lightweight strategy advocated in MIR literature for incorporating psychoacoustic bias without full auditory modeling, as discussed by Schedl et al. (2014) [8].

Layer 2 - Explainable feature groups. Grouping low- and mid-level audio features into musically interpretable categories reflects a well-established direction in MIR aimed at improving explainability. Timbre-related descriptors such as MFCCs and RMS energy are known to capture perceptual sound color and instrumentation, building on foundational perceptual studies by Grey [9] and genre classification work by Tzanetakis and Cook [10]. Spectral features such as centroid, bandwidth, and rolloff further describe brightness and spectral spread, complementing timbral descriptors as shown by Ellis [11]. Harmonic and chroma-based representations, originating from Fujishima's pitch-class profiles [12], have proven effective for capturing tonal structure and melodic similarity, with robustness improvements demonstrated by Gómez and formalized in Müller's music processing framework [13]. Rhythm and tempo features, rooted in Scheirer's beat-tracking work, form an independent descriptive dimension shown to be crucial for genre and style discrimination [14]. Organizing features into these groups aligns with MIR research emphasizing musically meaningful dimensions over isolated technical descriptors.

Layer 3 – UMAP-Based Similarity Embeddings. Dimensionality reduction techniques are commonly used in MIR to visualize and explore high-dimensional audio feature spaces. UMAP, introduced by McInnes et al. [15], has gained prominence due to its ability to preserve both local neighborhood structure and global manifold geometry, outperforming earlier methods such as PCA and t-SNE in many perceptual tasks. Survey work by Ghojogh et al. [16] highlights UMAP's suitability for complex, nonlinear data such as audio features. In music-oriented applications, UMAP has been successfully applied to visualize similarity relationships in learned and handcrafted audio embeddings, as demonstrated by Jiale and Yang (2020) [17] as well as Tovstogan et al. (2022) [18]. Computing separate embeddings for distinct feature groups follows recent visualization practices in MIR, enabling interpretable exploration of similarity along different musical dimensions while maintaining perceptual coherence.

Layer 4 - Query by Humming (QbH) has been extensively studied as an alternative to text-based music retrieval, addressing situations where users recall melodies but lack metadata. Early QbH systems relied on symbolic pitch representations, which proved sensitive to tuning errors and temporal variation. Salamon et al. (2014) [19] document the shift toward audio-based melodic representations that improve robustness in real-world conditions. Embedding-based approaches, as discussed by Humphrey et al. [20], enable fixed-length representations that support scalable similarity computation and tolerate expressive variability in hummed queries. Mel-spectral representations, in particular, have been shown to offer a perceptually aligned and noise-robust basis for melodic similarity, making them a common baseline in modern QbH systems. This line of work motivates the use of mel-based embeddings and distance-based retrieval as a practical and effective solution for melody-driven music search.

III. Methodology

A. Feature Extraction

Feature extraction is the process of computing a numerical representation that can be used to characterize a segment of audio. This numerical representation is called the feature vector, and is subsequently used as the fundamental building block of various types of analysis algorithms [5]. Features were selected based on the csv files provided in the Kaggle reupload of the GTZAN dataset. There were 60 columns in total with 57 notable audio features which can be categorized into chroma features, RMS energy, spectral centroids and bandwidth, spectral rolloff, zero-crossing, harmonic-percussive components, tempo/rhythm, and MFCCs. The existing csv files were not used in this experiment but they acted as a reference to feature extraction and all the manually features extracted for this experiment can also be found in the reuploaded version of the GTZAN dataset.

Chroma, better known as Pitch-Class Profile (PCP), is a low-level feature. PCP is a twelve-dimensional vector representing the intensities of the twelve pitch classes. In explainable terms, chroma features describe how much energy the signal contains for each musical class (frequencies of the same note name, e.g. note C, are grouped together) ignoring octave height. Chroma mean indicates the tonal center (key) and variance says how much the harmony changes over time.

Root Mean Square (RMS) energy is a well-known method in audio signal processing for estimating audio signal power and perceptual loudness. Higher RMS values (loudness) indicate louder segments while lower values indicate quieter ones. Mean RMS represents the global loudness, while variance reflects the dynamic contrast (flat vs. expressive dynamics).

Spectral centroid is the frequency-weighted average of the spectrum. It defines where the "center of mass" of the spectral energy lies. If more energy is concentrated at high frequencies,

the centroid shifts upward; if energy is concentrated at low frequencies, it shifts downward. Mean centroid captures overall brightness, while its variance captures how much brightness changes over time.

Spectral bandwidth measures how spread out the frequencies are around the centroid. It captures whether energy is tightly clustered or broadly distributed across frequencies. Narrow bandwidth represents a focused, tonal sound, whereas a wide bandwidth represents complex, noisy, or rich sound. Intuitively it asks how wide is the frequency range that actually matters. Mean bandwidth reflects overall spectral complexity and variance reflects timbral instability.

Rolloff frequency is the point below which a fixed percentage (typically 85%) of the total spectral energy is contained. It defines up to what frequency does most of the sound's energy extend. Mean rolloff indicates how much high-frequency content the signal usually has and its variance indicates how much that high-frequency presence varies.

Zero-crossing rate (ZCR) counts how often the waveform crosses the zero-amplitude axis (how rapidly it oscillates) within a frame. High-frequency or noisy signals cross zero often; low-frequency or smooth signals do not. Mean ZCR distinguishes noisy vs tonal content, variance captures transient or percussive activity.

Harmonic–Percussive Source Separation (HPSS): harmonic sounds represent the horizontal spectral continuity, percussive sounds represent vertical spectral continuity. *Harmonic component* isolates parts of the signal that are stable in frequency over time, such as sustained notes or pitched instruments. Mean harmonic energy indicates how tonal the track is, variance reflects melodic activity or sustain variation. It talks about how much of the sound is pitched and sustained. *Percussive component* captures short, broadband, time-localized events, such as drum hits or attacks. Mean percussive energy reflects rhythmic intensity, variance reflects rhythmic complexity and variation.

Tempo estimates the dominant periodicity of rhythmic events in the audio. This is done by detecting regular patterns in the onset strength over time. The strongest periodic pattern is converted into beats per minute (BPM).

Mel-Frequency Cepstral Coefficients (MFCCs) describe the overall shape of the spectral envelope on a perceptually motivated (mel) frequency scale. Instead of capturing exact frequencies, MFCCs capture how energy is distributed across low-to-high frequencies, in a way aligned with human hearing. Mean MFCCs describe the average timbral identity, variance captures articulation and expressive change.

B. Layer 1 - A-Weighted Pschoacoustic Features

The human ear is most sensitive to frequencies between 500 Hz and 8 kHz and responds less to very low-pitch or high-pitch noises. The frequency weightings used on a modern sound level meter are A-, C-, and Z-weightings. The most common weighting used in noise measurement is A-weighting which is the first layer to the music similarity algorithm this experiment proposes. A-weighting is a frequency-dependent weighting function derived from the equal-loudness contours, originally reported by Fletcher and Munson (1933), designed to approximate the sensitivity of human hearing at moderate sound pressure levels (SPLs) by reducing very low and high frequencies and emphasizing the mid-frequency range (approximately 2-5 kHz) where human auditory perception is most sensitive [6]. This perceptual model is standardized in IEC 61672 and enables spectral representations to reflect perceived rather than physical loudness.

In this experiment, A-weighting is applied at the feature level rather than to raw audio or Mel spectrograms. The input consists of pre-extracted Mel-band energy features stored in CSV format, where each Mel band represents a perceptually motivated frequency range. The center frequency of each band is used to evaluate the A-weighting curve, and the resulting weighting factors are applied directly to the corresponding Mel-band energies. Frequencies to which human hearing is less sensitive are attenuated, while perceptually salient mid-frequency bands are preserved, producing A-weighted feature vectors for downstream similarity computation [2] [7].

This feature-level approach provides a lightweight and interpretable perceptual approximation without performing signal-level processing or modeling nonlinear auditory effects such as masking. By biasing similarity measures toward perceptually relevant spectral regions, the method improves alignment with human judgments of timbre while maintaining computational efficiency and compatibility with CSV-based MIR workflows [8].

C. Layer 2 - Explainable Feature Groups

While A-weighting in layer 1 models how the human ear perceives relative loudness through A-weighted representations that approximate human loudness sensitivity, it remains focused on low-level auditory perception rather than musically interpretable structure. Layer 2 builds on this foundation by reorganizing mid-level audio descriptors into four perceptually and musically meaningful feature groups, thus shifting the system from perceptual weighting toward structurally interpretable musical dimensions.

Spectral and timbre features describe a sound's acoustic energy across different frequencies. They serve as a useful approximation for timbre—the perceptual quality that differentiates sounds with the same pitch and loudness [9]. In this experiment, **timbre** features include RMS energy and MFCCs, described by their mean and variance. They capture a sound's overall energy, spectral shape, and articulation. **Spectral** features consist of spectral centroid, bandwidth, and rolloff, also represented by their mean and variance. These quantify a sound's brightness, its spread across frequencies, and the distribution of high-frequency energy. Collectively, these descriptors capture characteristics such as instrumentation, production style, and timbral texture, which are valuable for tasks like genre classification and audio similarity [10]. Previous MIR research shows these features represent information that is separate from a song's harmony, providing a useful complement to pitch-based data [11]. **Chroma (harmony)**

features represent the distribution of spectral energy across the twelve pitch classes while collapsing information across octaves [12]. In this experiment, the chroma feature group contains statistics derived from a short-time Fourier transform (STFT mean and variance) as well as mean and variance descriptors derived from harmonic–percussive source separation. These features focus on tonal and harmonic information while remaining resistant to variations in timbre and pitch register. This makes them particularly effective for tasks like key detection and melodic similarity [13]. Enhanced harmonic representations focusing on stable spectral peaks have been shown to improve robustness in polyphonic and real-world recordings [21]. **Tempo (rhythm)** features describe how musical events are organized in time by capturing patterns in pulse and meter [14]. This group includes a global tempo estimate and statistics for zero-crossing rate (mean and variance). These features describe higher-level temporal patterns that simpler spectral or harmonic descriptors cannot directly access. They have been shown to form an independent dimension of musical description in MIR systems [10]. As a result, rhythm-based representations are particularly informative for distinguishing musical styles and explaining similarities driven by temporal structure.

This new layer organizes audio features into intuitive groups like timbre, harmony, and rhythm, rather than using individual technical metrics. This makes it possible to explain similarity (e.g. saying that two songs sound alike because of their shared harmony). These groups reflect well-established musical qualities and provide a clearer, more interpretable foundation for comparing music. The next layer will use these grouped features to calculate similarity, allowing control over how much each musical aspect influences the final result.

D. Layer 3 - UMAP-Based Similarity Embeddings

After computing feature-specific similarities in the second layer, layer 3 applies *Uniform Manifold Approximation and Projection (UMAP)* to produce low-dimensional embeddings of the high-dimensional audio feature space. Each feature group (timbre, spectral, rhythm/tempo, chroma/harmony, and A-weighted perceptual features) is embedded separately, producing multiple 2D visualizations that capture distinct aspects of musical similarity. UMAP is a nonlinear dimensionality reduction technique that constructs a continuous manifold from high-dimensional data while preserving both local neighborhoods and global structure [15] [16]. It is particularly suited for audio and music information retrieval tasks, where features such as MFCCs, spectral descriptors, and chroma vectors are high-dimensional and perceptual relationships between tracks are obscured [18]. Computing one UMAP per feature group preserves the interpretability established in Layers 1–2. Each feature group represents a semantically distinct aspect of musical content:

- **Timbre UMAP**: Encodes instrument textures and sonic color, primarily derived from MFCCs and RMS features. Clusters in this space reveal similarities in timbral characteristics across tracks.
- **Spectral UMAP**: Captures brightness and frequency distribution patterns using spectral centroid, bandwidth, and rolloff. Tracks close in this embedding share similar spectral envelopes.
- **Rhythm/Tempo UMAP**: Represents rhythmic energy and tempo similarity, informed by zero-crossing rates and beat-tracking. This allows the visualization of tempo- and rhythm-related clusters.
- **Chroma/Harmony UMAP**: Emphasizes tonal and harmonic relationships, organizing tracks according to chord structures and pitch content.
- **A-weighted Mel UMAP**: Integrates perceptually weighted loudness and energy features, reflecting the human auditory response in the clustering structure.

Each embedding provides a musically interpretable 2D space where proximity corresponds to similarity within that feature group. Comparing embeddings across groups enables multi-dimensional exploration: a track may cluster by timbre in one embedding while aligning differently in rhythm or harmony space, revealing nuanced relationships. This approach preserves explainability and interpretability from layers 1 and 2 [17], supporting both analytical insight and interactive exploration of complex music collections [18] [16].

E. Layer 4 - Query by Humming

Query by Humming (QbH) is a MIR task in which users retrieve musical works by vocally imitating a melody rather than using textual metadata. Hummed queries differ significantly from studio recordings in timbre, pitch stability, tempo, and recording conditions. These differences require audio representations that emphasize perceptually relevant musical content while remaining robust to noise and performance variability. Early QbH systems relied on symbolic representations such as pitch contours or note sequences, but these approaches are sensitive to pitch extraction errors and temporal misalignment [19]. To address these limitations, recent systems adopt embedding-based retrieval, where both queries and reference tracks are represented as fixed-length vectors and compared using standard distance metrics [20].

In this project, QbH is implemented using mel-spectral embeddings with statistical pooling. Each audio signal is converted to a monophonic waveform and transformed into a mel spectrogram using a perceptually motivated frequency scale. Spectrograms are converted to a logarithmic amplitude representation and then center-cropped or zero-padded to a fixed number of time frames to ensure comparability across inputs. Per-frequency-band z-score normalization is applied to reduce loudness bias and inter-recording variability. Temporal information is summarized by computing the mean and standard deviation of each mel band over time, producing a fixed-length embedding. Both reference tracks and hummed queries are processed using the same pipeline. Retrieval is performed by computing cosine distances between embeddings, yielding a temporally invariant and noise-robust baseline QbH system.

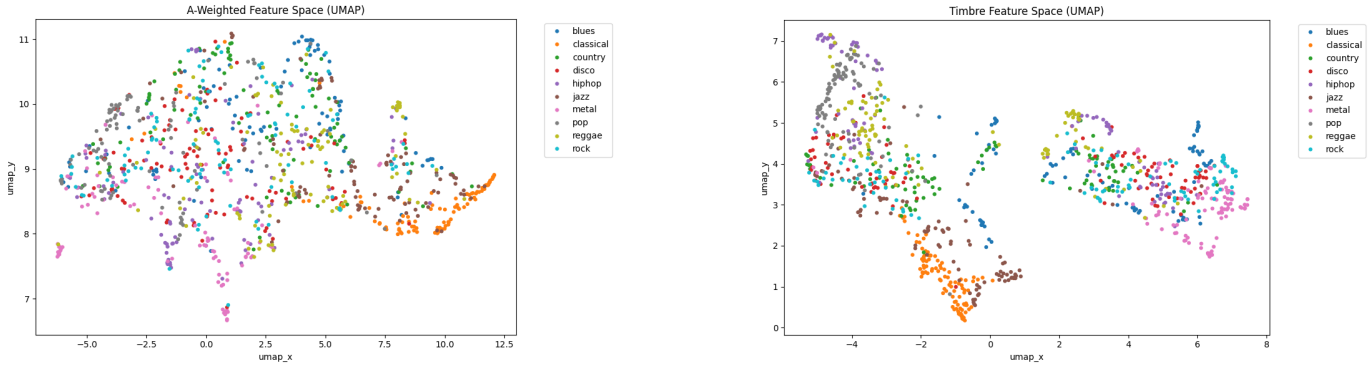

Fig. 1. Comparison of A-weighted and Timbre UMAP visualizations

F. User Interface

The user interface was designed to make music similarity exploration intuitive and interactive, even for users without a technical background. The application is organized into three main panels. The left panel contains a music player with a playlist of all available tracks from the GTZAN dataset, displaying the corresponding album cover, track name, and playback timeline to provide immediate auditory context. The center panel focuses on similarity exploration: it presents a list of tracks that are most similar to the currently selected song or to a user-provided humming query (top 10 most similar), allowing users to audition recommended tracks directly without disrupting the main playback. The right panel visualizes multiple similarity spaces using UMAP projections, where each map represents a different group of audio features and shows how songs cluster based on perceptual or musical characteristics. Together, these components create a cohesive interface that combines listening, comparison, and visualization, enabling users to explore music similarity both audibly and visually in a seamless workflow.

IV. Results

For a representative example, the track *metal00027* was selected as the query for the layered similarity search. The system returned the top five most similar tracks based on the aggregated distance across perceptual loudness (Layer 1), timbral features, spectral features, rhythmic characteristics, and chroma content (Layers 2–3). As expected, the closest matches were predominantly drawn from the same genre, with *metal00082* and *metal00072* ranked highest, indicating strong overall similarity. The total distance reflects a weighted combination of individual layer distances, where smaller values correspond to greater similarity. In this example, timbral similarity contributed most strongly to the ranking, reflecting shared sound texture and instrumentation typical of metal recordings, while spectral and rhythmic–temporal distances were consistently low across the top results, indicating comparable energy distribution and tempo-related characteristics. Notably, tracks from adjacent genres such as rock (*rock00087*) and even country (*country00005*) appeared among the top results, illustrating that the system captures perceptual similarities beyond strict genre labels when multiple feature layers are jointly considered.

The UMAP results include five separate two-dimensional projections, one for each feature set. Figure 1 shows two of these UMAPs. In each plot, the horizontal (x) and vertical (y) axes represent different acoustic or perceptual characteristics.

The A-weighted UMAP captures perceived loudness and frequency balance, where the x-axis distinguishes tracks with more high-frequency energy (right) from those with more low-frequency energy (left), and the y-axis reflects perceptual loudness, from quieter, less dynamic tracks (bottom) to louder, more energetic ones (top). The timbre UMAP organizes tracks by their sound texture, with the x-axis ranging from bright, sharp timbres (right) to dark, soft timbres (left), and the y-axis separating percussive, transient-rich tracks (bottom) from smooth, sustained textures (top).

Three of these UMAPs are not shown in the figure but are provided in the supplementary notebook `results.ipynb`[2].

The chroma/harmony UMAP reveals tonal structure: the x-axis contrasts brighter, major-like harmony (right) with darker, minor-like harmony (left), while the y-axis distinguishes simple harmonic progressions (bottom) from complex chordal structures (top). The rhythm/tempo UMAP emphasizes temporal patterns, arranging tracks from slower (left) to faster (right) tempos along the x-axis, and separating simple, steady rhythms (bottom) from complex, irregular patterns (top) along the y-axis. The spectral UMAP focuses on frequency distribution, where the x-axis separates high-frequency dominant tracks (right) from low-frequency dominant ones (left), and the y-axis represents spectral stability, from steady (bottom) to fluctuating (top).

The QbH results form a separate retrieval layer and are not combined with the multi-layer similarity model. For a hummed query, the system extracts a mel-spectrogram representation and compares it directly to a dedicated QbH feature dataset. The top ten matches all achieve very high cosine similarity values (above 0.98), with most results belonging to the pop and reggae genres; importantly, the second most similar track

[2]https://github.com/kanitafro/music-similarity

(pop00032) is the intended target of the humming query. This behavior reflects the melody-driven nature of QbH, which focuses on pitch movement and timing rather than timbre or production details. Together, these results show that while the layered similarity system models overall perceptual similarity, QbH provides an effective way to retrieve music based solely on vocal melody input.

Table 1. QbH results for *queries/humming2.wav*

track_id	genre	similarity
pop00064	pop	0.983434
pop00032	pop	0.983204
hiphop00037	hiphop	0.983119
pop00055	pop	0.982884
reggae00088	reggae	0.981342
pop00097	pop	0.981117
pop00048	pop	0.980685
reggae00072	reggae	0.980614
pop00077	pop	0.980483
reggae00087	reggae	0.980379

V. Conclusions and Future Work

This project presented a multi-layer music similarity and retrieval framework that integrates perceptual, spectral, rhythmic, harmonic, and melodic representations to model musical similarity in a structured and interpretable manner. By combining feature-specific similarity layers with UMAP-based visualizations, the system delivers both quantitative similarity scores and intuitive qualitative insight into relationships between tracks across multiple musical dimensions. The results show that meaningful similarities frequently emerge beyond genre boundaries, reflecting shared perceptual and structural characteristics rather than categorical labels alone.

The modular design enables flexible analysis and transparent interpretation, allowing individual feature groups to be examined independently while contributing to a unified similarity measure. The separation of query-by-humming from audio-based similarity further highlights the distinct role of melodic contour in music retrieval. Together, these components demonstrate how layered, perceptually motivated representations can support explainable and user-oriented music similarity systems and provide a strong foundation for future extensions such as feature weighting, personalization, or interactive exploration.

Future work will expand this system beyond music similarity to also classify genres. Using the same audio features, we can train a simple model on the GTZAN dataset to predict which of its ten genres a track belongs to. This model would output a probability for each genre, allowing it to make single predictions and also show how certain it is. The current layered feature design is flexible and can be adapted to other datasets. By keeping the feature extraction separate from the dataset labels, the same system could support both similarity search and genre classification at the same time. This would make it a reusable tool for analyzing music, where perceptual, spectral, and rhythmic features provide understandable inputs for multiple tasks across different collections of audio.

References

[1] J. Futrelle and J. S. Downie, "Interdisciplinary research issues in music information retrieval: ISMIR 2000–2002," *Journal of New Music Research*, vol. 32, no. 2, pp. 121–131, 2003.

[2] P. Knees and M. Schedl, *Music similarity and retrieval: an introduction to audio-and web-based strategies*. Springer, 2016, vol. 36.

[3] T. George, E. Georg, and C. Perry, "Automatic musical genre classification of audio signals," in *Proceedings of the 2nd international symposium on music information retrieval, Indiana*, vol. 144, 2001.

[4] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[5] G. Tzanetakis and P. Cook, "Music analysis and retrieval systems for audio signals," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 12, pp. 1077–1083, 2004.

[6] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *Bell System Technical Journal*, vol. 12, no. 4, pp. 377–430, 1933.

[7] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011.

[8] M. Schedl, E. Gómez, J. Urbano *et al.*, "Music information retrieval: Recent developments and applications," *Foundations and Trends® in Information Retrieval*, vol. 8, no. 2-3, pp. 127–261, 2014.

[9] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *the Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, 1977.

[10] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[11] D. P. Ellis, "Classifying music audio with timbral and chroma features," 2007.

[12] T. Fujishima, "Realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music," in *International Conference on Mathematics and Computing*, 1999.

[13] M. Müller, *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015, vol. 5.

[14] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[15] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[16] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Uniform manifold approximation and projection (umap) and its variants: tutorial and survey," *arXiv preprint arXiv:2109.02508*, 2021.

[17] Y. Jiale and Z. Ying, "Visualization method of sound effect retrieval based on UMAP," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1. IEEE, 2020, pp. 2216–2220.

[18] P. Tovstogan, X. Serra, and D. Bogdanov, "Visualization of deep audio embeddings for music exploration and rediscovery," *Proceedings of the SMC 2022 Music technology and design*, pp. 493–500, 2022.

[19] J. Salamon, E. Gómez, D. P. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[20] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Feature learning and deep architectures: New directions for music informatics," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 461–481, 2013.

[21] E. Gómez, "Tonal description of polyphonic audio for music content processing," *INFORMS Journal on Computing*, vol. 18, no. 3, pp. 294–304, 2006.