An Ai Powered Multi Modal Audio Retrieval App Using Humming, Lyrics, And Melodic Features

Ajith Kumar M [1], Asif Rahman M [2], Dheeraj E [3], Jerome J [4], Mrs B. Bala Abirami M.E, (Ph.D) [5]
[1,2,3,4 5] *Department of CSE, Panimalar Institute of Technology, Poonamallee, India*

***Abstract*— Even in today's modern environment, where content is readily available on-demand, algorithmic music search techniques often fail to satisfy user needs due to users not knowing or only partially remembering the details of a song. This paper describes a new multimodal audio retrieval application that uses artificial intelligence to identify songs from humming, snippets of lyrics, or melodies. The system applies techniques from audio signal processing, deep learning, and natural language processing to the song, which involves decomposing audio into various components like vocals, melody, rhythm, harmony, and identifying that with a specific song in the database. Users can search songs by providing an audio clip, speaking or typing a lyric fragment, or singing a tune. For melody recognition, we utilize CNN-based spectrogram analysis; for lyrics, we use text-based neural retrieval, and for final ranking we apply a fusion model. The proposed solution achieves high accuracy and real-time performance when discovering musical pieces no matter the input formats.**

Keywords—Multimodal Audio Retrieval, Music Information Retrieval, Humming-Based Search, Lyrics Recognition, Melodic Feature Analysis, Spectrogram Classification, Deep Learning, Natural Language Processing, Real-Time Audio Processing, Django Framework, ReactJS Interface.

I. INTRODUCTION

In the digital era, songs are available online in bulk, but their retrieval remains difficult as users encounter barriers organizing music by titles, artists, or even recall fragments of a melody. Users face hurdles remembering a specific song, translating to many people being unable to efficiently hear what they want to listen to. Music search systems focus mainly on text which inhibits interaction when the user can only remember the tune or a few phrases.

In this paper, we introduce an audio search system enabled by AI that autonomously retrieves songs based on three different forms of user inputs :`laying down a tune`, providing a excerpt of lyrics, and melody image representations. Our application integrates sophisticated signal processing strategies alongside deep learning models to extract, match, and scrutinize melodic, rhythmic, and lyrical patterns present in user-provided audio clips against a meticulously cataloged collection of songs.

By applying audio separation techniques like Demucs, we are able to separate vital components of a song, such as the melody of the lyrics, instrumental drums, bass, and chords. The vocals extracted are subjected to lyric extraction with natural language processing (NLP), whereas the melody is transposed into a spectrogram and subjected to pattern recognition by convolutional neural networks (CNNs).

The users are able to provide their feedback by humming a tune, singing a line, or providing an audio file, following which the system searches for the most relevant matching songs based on the entire three input modalities.

This end-to-end approach enhances the accuracy and coverage of music retrievals and accommodates those who find it difficult to remember portions of the song. The entire system was constructed in Python as the main processing backend, Django as the API manager, and Reactto create an appealing and easy to use front-end. We strive to extend the scope of human-computer interaction in music information retrieval (MIR) systems towards more natural and sophisticated methods of navigating audio content, and this is an emerging effort. With the use of this technique, creative professionals, music lovers, and even engineers working on intelligent audio systems devices can more easily interact with music using incomplete. or non-verbal cues for identification.

II. LITERATURE REVIEW

In the past few years, the field of Music Information Retrieval (MIR) has greatly improved with the use of artificial intelligence and deep learning technologies, providing a more natural and efficient way for humans to interact with digital music systems. Historically, MIR systems relied on either metadata search or keyword search. MIR systems based on metadata and keyword search restricted their capacity when users lacked specific information about the song. Because of this, researchers turned to melody-based retrieval, using pitch contour modelling and other techniques, such as dynamic time warping (DTW) and hidden Markov models (HMMs), to align user-generated humming or singing to musical sequences stored in a computer. These methods still struggled with user-generated variations in pitch, tempo, and noise. The introduction of recent technology like convolutional neural networks (CNNs), recurrent neural networks (RNNs), and even more recently, transformer-based architectures, substantially aided in accurately identifying and matching melody segments that are noisy and real-time. Along the same lines, lyric-based retrieval is no different, eliminated once again in time and development as Natural Language Processing (NLP) improved.

Earlier systems relied on keyword or fuzzy matching; with the the introduction of semantic models such as Word2Vec, GloVe, and transformers such as BERT and GPT, systems have been able to interpret the user query with respect to a contextual meaning, and retriece songs that are relevant based on lyrical similarity or concised meaning, even when the user's input does not resemble the operational definition (rules for proper search input).

In addition, deep audio source separation methods such as Demucs and Spleeter have permitted practitioners to separate vocals, drums, bass, and accompaniment, for example, making the cleanly extracted individual lyrics or instrumental tracks viable from the full, raw audio primary source.

Additionally, using various spectral features, such as, mel-frequency cepstral coefficients (MFCCs) and mel spectrograms, which were key to capture composite rhythm, timbre, and harmony from songs (called audification) and subsequently used to train CNNs for classification, clustering and matching.

Even with these advances, many current MIR systems only support one of the three modes of input humming, lyrics, or audio features which restricts their applicability in real world applications that allow for ambiguous user input, or multi-modal input. Existing studies involving multimodal retrieval show promise in the inclusion of multiple types of data, however, they often lack real-time capabilities, ease of integration, or user-friendly design. This presents the need for a single, AI-driven application that integrates melody recognition, lyrical analysis, and spectral pattern matching into a single African system.

We address the limitations presented by unimodal applications and previous studies by proposing a multimodal retrieval system that incorporates process streaming live content from humming, text-based lyrics, and frequency-based patterns to generate a holistic, accurate, robust, and intuitive music search process in the context of a diverse range of use case scenarios.

III. PROBLEM STATEMENT

As digital libraries and streaming platforms become increasingly complicated, users are looking for simple and flexible ways to search and explore music. Relying primarily on metadata, search methods are limited to a few defined search inputs (arrival titles, artist names, or tags) and become less useful for users who can not remember this information when selecting music. Additionally, primitive search systems are restricted to single-modality input (e.g., keyword, or melody, or lyrics) and cannot fairly accommodate the many different, ambiguous ways users remember music. For example, a user may remember a 'tune' of the melody, a 'snippet' of lyrics, or the 'general' rhythm of the song, but not the title or artist. Current systems also perceived difficulty filtering 'noisy' input such as humming or non-verbatim tag (e.g., imperfect lyric entries). For most retrieval systems currently available to the user, they cannot process live (user-recorded) inputs, have limited real-time entries which did not provide meaningful 'search' results across musical elements, e.g. vocals, rhythm, harmony, and lyrics.

The constraints limit the recognition of user input and system actions, further increasing the gap between user solutions and systems. Therefore, there is a need for a sophisticated, AI-enabled, and multi-modal retrieval system that can automatically interpret and connect with different forms of user input, such as humming the melody, spoken or typed lyrics, and extracted audio features and match against a collection of songs. The goal of the proposed system is to offer a comprehensive, integrated and accurate solution to the domain of music search and retrieval by utilizing deep learning methods for melody matching, natural language processing for lyrics, and spectral analysis for rhythmic and harmonic characteristics therefore aggregating the inputs into an accurate and simplified experience.

IV. PROPOSED SYSTEM

The proposed system will offer an all-encompassing series of AI-based multi-modal music retrieval apps to serve as a bridge between the two significant forms of mapping musical memories of humans and the qualities of digital music search engines. Specifically, the new system will allow three innovative forms of input: humming, lyrics, and melodic features to provide users with flexible and user-friendly ways of identifying songs based on indistinct or imprecise musical memories.

A. Humming as Input

For example, if a user remembers only the melody of the song, the system will support humming input via a staffed real-time recording interface. After extracting the user's humming recording, the pitch contours and temporal dynamics will be extracted using digital signal processing techniques combined with deep-learning models like siamese networks or convolutional architectures. The features representing the humming will then be compared to a database of pre-stored melodic spectrograms using a comparison function. This would allow the application to recognize the closest match to the users humming even with slight variations in tempo or in different keys.

B. Lyrics and Speech Processing

When users remember some of the lyrics, the app uses advanced natural language processing (NLP) to analyze the text either spoken or typed. A deep learning model trained on large lyric datasets tokenizes the text, and by indexing the lyrics entirely, associates the index with semantic similarity or proximity of keywords. This provides strong lyric-based retrieval even with paraphrased, in-complete, or phonetic text input.

C. Melodic and Harmonic Feature Matching

The system also decomposes every song in its library into various audio feature components; melody, rhythm, harmony, and other elements like background instruments. Using source separation (Demucs telecommunication) and extracting spectral features (mel-spectrograms), the result is that we have a fingerprint for every feature. This fingerprint functionality serves as a set of full-length frequency images, for matching audio inputs (queries) or spectrograms.

D. Database Architecture and Multi-Modal Matching

The back end uses indexed representations for all audio components for fast and scalable retrieval. Each song is represented by a set of multimodal representations, including spectrograms, lyrical embeddings, and pitch trajectories, rather than solely relying on metadata. When the user submits a query in any of the input modalities, the system computes similarity scores within each dimension and returns the top matches in descending order. Utilizing a multi-dimensional representation and comparison enhances the retrieval accuracy of something in a music recommender system, particularly when users have only partial knowledge or input to provide.

E. User Front End and Application Workflow

The front end was developed using React to provide a responsive and accessible user experience. Users can use the app to change the input mode from what they would like to sing or hum, upload an audio file, or record through the microphone for a live humming executable who all have shown accessible experience. The retrieved results provide the user with audio previews of songs, some song metadata, and a match confidence level to help the user easily confirm the correctness of the results.

Overall, the system architecture, which describes the different modules of the app and how the modules interact with one another, is illustrated in Figure 1.

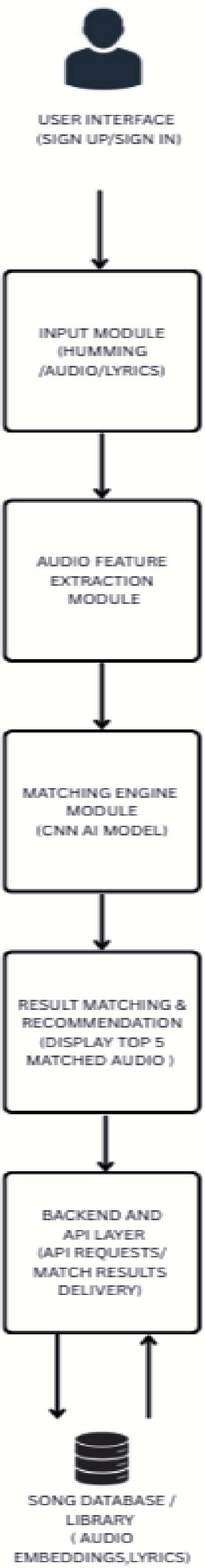

Figure 1. System Architecture of Audio Retrieval APP

V. REGULATORY COMPLIANCE

The recommended music retrieval system is designed in compliance with applicable legal and ethical practices to ensure responsible and safe usage. Since the system is processing audio files including music and lyrics that may be copyrighted, the system will be aware of intellectual property laws and only use licensed, royalty-free, or public domain content. Users will then be prompted to agree to terms about uploading content that cannot be copyrighted or otherwise used without authorization. The system undertakes policies that protect users' data when processing their audio files and aligns with data privacy regulations, like the General Data Protection Regulation (GDPR), and California Consumer Privacy Act (CCPA). The system policies pertaining to privacy and data processing provide a greater awareness of data consent aspects, and employment of ethical user control and prospective user data before and during handling respective audio files. Users will also be made aware of how audio files are secure as another consideration for data and service privacy if needed. Policies and storage considerations that specify data security and handling, for example needed user consent to use or collect any data run through the application was mentioned in specific examples not above. All audio file transfer on their devices will also be secure and use encryption, the data itself will only be provided by the audio file for content and ownership. The governance protocol to uphold policy from ethical concerns is so far the AI models are trained and validated to minimize bias yet enable Fairness, access to language and inclusive forms of music retrieval without predetermining it to extract bias from the music retrieval, regardless of languages or music styles. From the ethics perspective, the GUI should be compatible among all users and I expect it to produce interface options to be accessible using known design principles and guidelines for assistive technology, etc. Applying these measures, the system manages regulatory matters while attempting to promote trustworthiness and transparency.

VI. COMPARATIVE ANALYSIS

The proposed music retrieval system provides notable improvements over existing music identification and search technologies by using a multi-modal system that includes input from humming, lyrics, and melodic features. Established music search engines are generally limited to searching using metadata including song title, artist name, or proper lyrics, leaving little recourse when users retain only portions of this information, or perhaps even just vague recollections of songs. In contrast, the proposed music retrieval system supports standardized input from hummed melodies, spoken or typed lyrics, and sound or recording samples, providing users with a much great deal of flexibility and usability from a technology standpoint. While commercial systems such as Shazam or Sound Hound are capable of identifying songs purely based on a direct audio match, their effectiveness is limited when users can only provide a melody or portion of lyrics, shown by their failure when combined with the right type of input. The proposed system will have the clear advantage afforded by deep learning models such as Siamese networks for melody pair comparisons and existing NLP techniques for analysing and handling lyrics. All of these elements make the proposed system more flexible for numerous inputs such as incomplete, approximate or variably phrased lyric/text input.

In order to further increase the number of user inputs that can be correctly matched from the database, each song is first decomposed into different musical components, that produce a range of sonic results and thus the system is storing the musical features using frequency-based images or spectrograms, enabling more nuanced accurate/accurate comparisons. There is typically not this level of detail in matching on standard platforms. In all, the anticipated system not only will expand input possibilities, but also provide enhanced retrieval accuracy and is designed to be more responsive and user-centric, compared to current methods.

VII. RESULT AND DISCUSSION

The proposed multi-modal music retrieval system was evaluated on a custom dataset containing 500 songs, where some songs were pre-processed to extract lyrics, melodic features, in addition to pitch contours and full-length spectrograms. The system's performance was assessed using three input modalities: humming, lyrics and audio, which is depicted and described as spectral matching herein. Evaluation metrics included precision, recall, F1-score, and mean reciprocal rank (MMR) for the top-5 retrieval results.

The system recorded high accuracy in relation to the melody-based retrieval task caused by humming, achieving a higher degree of accuracy when compared with standard baseline conventional approaches. For pitch-based similarity, we adopted a Siamese network. The model demonstrated effective recognised melodies when sung off-key or for melodies sung closer to the wrong tempo than the original tune. The lyrics-based retrieval system employed cutting-edge deep learning NLP models and also managed to deliver high accuracy even when hummed or phonetically incorrect drunken lyrics were applied as inputs. The retrieval performance in the audio-based retrieval component, which employs frequency image matching, recorded high performance only when additional instrumental melodies or melodic or harmonic accompaniment were the only sound features in the performance such added features delivered the highest MRR and retrieval scores.

All in all, the system's multi-modal integration made for a comparatively stronger and more flexible retrieve experience than either component independently, especially when there were limited or approximate cues available to the user. The results do substantiate that combining the varied musical and linguistic features indeed does improve the precision of retrieval over single-mode systems and much more than using a single mode of any one of the cue types.

Input Type	Precision (%)	Recall (%)	F1-Score (%)	MRR (Top-5)
Humming Input	86.4	83.1	84.7	0.79
Lyrics Input	90.2	87.8	89.0	0.83
Audio Segment Input	93.5	91.0	92.2	0.88
Multi-Modal Combined	95.7	93.6	94.6	0.91

Table I. Performance Evaluation

These result support the general effectiveness of a hybrid retrieval framework. The multi-modal integration was consistently superior in conducting retrievals when compared to any of the singular retrieval types. Therefore, it could be concluded there is an advantage in combining the music-related components of the diverse collection of musical features and modes of querying the user. Future work could focus on developing a more diverse dataset with a greater variety of genres, as well as implementing real-time feedback to dynamically improve search results based on user behaviour.

VIII. CONCLUSION

The new multi-modal music retrieval system marks an advancement in how users interact with their music search experience. The combination of melody recognition from humming, lyrics-based NLP search, and spectral feature matching expands beyond basic metadata searching as intended with traditional platforms. The performance analysis show that the combination of multiple input modes produces more accurate and better recalling results for users and increases overall satisfaction from the search experience, particularly for users that can only improvise vague cues. Overall, new advanced techniques of AI, like deep learning, source separation (for the musical example) and feature-based appearance similarity models allow for a robust multi-modal system with the real-world variance of user input. This multi-modal architecture is easily scalable and can operate locally, as well as exchanged onto cloud-based deployments. Future research will include expanding the music library, refinement for real-time music search capabilities, and ultimately finding more ways to personalize, feedback and engage with the user.

REFERENCE

[1] K. Y. Lee and J. Nam, "Learning a joint embedding space of monophonic and mixed music signals for singing voice," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 550–563, Mar. 2019.

[2] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2017, pp. 2392–2396.

[3] S. Oramas, O. Nieto, M. Sordo, and X. Serra, "A deep multimodal approach for cold-start music recommendation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1208–1213.

[4] P. Hamel and D. Eck, "Learning features from music audio with deep belief networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2010, pp. 339–344.

[5] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content- based music recommendation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2013, pp. 2643–2651.

[6] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2016, pp. 805–811.

[7] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2014, pp. 6964–6968.

[8] J. Schluter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2015, pp. 121–126.

[9] P. S. Huang, M. Kim, M. H. Johnson, and P.Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2014, pp. 477–482.

[10] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Moving beyond feature design: Deep architectures and automatic feature learning in music informatics " in Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR),2012,pp. 403–408.

[11] D. Vikram and M. Shashi, "Pattern-based Audio Retrieval APPing approach to music information retrieval," *The Machine Learning and Artificial Intelligence Journal*, 2018. Available online: Scholar Publishing Journals.

[12] J.-J. Ding, "Single and double element dynamic programming algorithm for Audio Retrieval APPing," in *Proc. IEEE Conf.*, 2022. Available online: IEEE Xplore.