IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 2, Ser. 1 (Mar. – Apr. 2026), PP 90-97
www.iosrjournals.org

Hum2Song: An AI-Driven System For Converting Hummed Melodies Into Complete Musical Compositions

Chirag & Parshant Kumar Gupta

Abstract:
Recent advancements in artificial intelligence have significantly influenced creative fields such as music composition. However, many individuals who possess musical ideas often lack the technical knowledge required to convert those ideas into structured musical compositions using professional tools. This paper presents Hum2Song, an AI-powered web-based system designed to transform hummed melodies into complete musical compositions. The proposed system captures a short vocal melody from the user through a browser-based interface and processes the audio using machine learning techniques to extract musical features such as pitch, timing, and note duration.
The detected melody is converted into a symbolic music representation using MIDI (Musical Instrument Digital Interface), enabling further analysis and processing. A generative AI model then produces harmonically consistent accompaniment tracks, including chords, basslines, and rhythmic patterns, based on the extracted melody. The final multi-instrument composition is synthesized into audio using digital instrument libraries, allowing users to play or download the generated music.
The system is implemented as a client-side web application, leveraging modern browser technologies such as TensorFlow.js and Magenta.js, enabling machine learning models to run directly within the user's browser without requiring server-side processing. This architecture improves accessibility, reduces latency, and ensures user privacy by keeping audio data local to the device.
To evaluate the effectiveness of the proposed system, experiments were conducted to measure note detection accuracy, processing time, and user satisfaction. Results demonstrate that the system can successfully convert hummed melodies into musically coherent compositions while maintaining efficient performance. The proposed approach highlights the potential of AI-assisted music generation systems to democratize music creation and enable intuitive human–AI collaboration in the creative process.

Date of Submission: 06-04-2026 Date of Acceptance: 16-04-2026

I. Introduction

Music composition has traditionally been a complex process that requires knowledge of music theory, instrumentation, and specialized software tools such as digital audio workstations (DAWs). While professional musicians and producers use these tools to transform musical ideas into complete compositions, many individuals who possess creative musical ideas often lack the technical expertise required to translate those ideas into structured music. In many cases, people naturally express musical thoughts by humming or singing simple melodies. However, converting these spontaneous vocal expressions into full musical arrangements typically requires significant training and experience in music production.

In recent years, advancements in artificial intelligence and machine learning have opened new possibilities in the field of music generation and computational creativity. AI-based systems are increasingly being used to assist in creative tasks such as melody generation, harmonic composition, and automatic music production. Several modern AI tools and platforms have demonstrated the ability to generate music from textual descriptions, predefined templates, or stylistic inputs. These systems utilize deep learning models, including transformer architectures and recurrent neural networks, to learn musical patterns from large datasets and generate coherent musical sequences.

Despite these advancements, many existing AI music generation tools primarily focus on generating music from text prompts or preset parameters. While these approaches are powerful, they often overlook one of the most natural forms of musical expression: the human voice. Musicians frequently hum melodies while brainstorming musical ideas, making vocal input a natural and intuitive interface for music creation. However, accurately converting humming or singing into structured musical representations remains a challenging task due to factors such as pitch variations, background noise, and irregular timing in human vocalization.

To address this challenge, this paper presents Hum2Song, an AI-powered system designed to transform hummed melodies into complete musical compositions. The proposed system captures vocal input from a user through a web-based interface and applies machine learning techniques to analyze the audio signal and extract musical features such as pitch, timing, and note duration. The detected melody is then converted into a symbolic music representation using MIDI, which allows the melody to be digitally processed and extended.

Once the melody has been transcribed into MIDI format, the system utilizes generative music models to produce harmonic accompaniment, including chords, basslines, and rhythmic patterns that complement the original melody. The generated musical elements are combined to produce a multi-instrument composition that can be played directly in the browser or exported for further editing. By enabling users to generate music simply by humming a melody, the system lowers the barrier to music creation and encourages collaborative interaction between human creativity and artificial intelligence.

II. Literature Review

Foundations of Artificial Intelligence in Music Generation

Early research in computational music generation focused on rule-based systems that encoded music theory into algorithmic frameworks. However, with the growth of machine learning, researchers began exploring data-driven methods capable of learning musical patterns directly from large datasets. One influential contribution is the work of Agostinelli et al. (2023), who introduced MusicLM, a generative model capable of producing high-quality musical compositions from textual descriptions. The study demonstrated that deep learning models can capture hierarchical musical structures and generate coherent compositions while preserving rhythm, harmony, and style.

Similarly, Copet et al. (2023) proposed MusicGen, a transformer-based architecture designed to generate music conditioned on textual prompts and musical attributes.

Their work showed that transformer models are capable of learning complex temporal relationships within musical sequences, allowing the system to generate musically consistent outputs. These foundational studies highlight the growing capability of artificial intelligence to generate structured musical compositions. However, most of these systems rely primarily on textual input rather than natural human vocal expressions, leaving a gap in systems that can directly interpret humming or singing as musical input.

Advances in Pitch Detection and Audio-to-MIDI Transcription

Accurate pitch detection is a fundamental component in systems that convert vocal melodies into structured musical representations. Early approaches relied on signal processing techniques such as the Fast Fourier Transform (FFT) and autocorrelation methods to estimate pitch from audio signals. While these techniques provided initial solutions for pitch estimation, they often struggled with noisy audio signals and the natural variability present in human humming.

Recent research has introduced machine learning-based pitch detection models that significantly improve transcription accuracy. CREPE, developed by Kim et al. (2018), employs a convolutional neural network to estimate pitch directly from raw audio waveforms. The model demonstrated improved accuracy compared to traditional signal processing techniques, particularly in complex audio environments. More recently, Basic Pitch, introduced by Spotify researchers, provides a lightweight neural network model specifically designed for audio-to-MIDI transcription. This model is capable of detecting note onsets, offsets, and pitch variations with high precision while maintaining computational efficiency, making it suitable for real-time and browser-based applications.

Generative Models for Symbolic Music Composition

Symbolic music generation focuses on producing music in structured formats such as MIDI, where musical events are represented as discrete notes, durations, and velocities. This representation allows machine learning models to analyze and generate music while preserving important musical relationships. One significant contribution in this area is MusicVAE, introduced by Roberts et al. (2018) as part of the Magenta research project. MusicVAE uses a variational autoencoder architecture to learn latent representations of musical sequences and generate stylistically consistent variations of melodies.

Transformer-based architectures have also gained prominence in symbolic music generation due to their ability to model long-range dependencies in sequential data. Recent research such as Midi-LLM (Yuan et al., 2025) explores adapting large language models for music generation by treating MIDI events as tokenized sequences similar to natural language. These models demonstrate that deep learning architectures can learn complex harmonic structures and generate coherent accompaniments when conditioned on a melody input.

Evaluation of AI-Based Music Generation Systems

Evaluating AI-generated music presents unique challenges due to the subjective nature of musical quality and listener perception. Researchers typically employ a combination of objective performance metrics and human evaluation studies to assess system effectiveness. Objective measures often include pitch detection accuracy, note onset detection accuracy, and computational performance, which provide quantitative insights into system reliability.

In addition to technical metrics, user studies are frequently conducted to evaluate the perceptual quality and musical coherence of generated compositions. Donahue et al. (2023) investigated the effectiveness of AI-generated musical accompaniments through controlled listening experiments, comparing outputs generated by AI systems with human-composed music. Their findings highlight the importance of combining algorithmic evaluation with human feedback when assessing creative AI systems. Such evaluation frameworks help determine whether AI-generated music is both technically accurate and aesthetically meaningful for listeners.

III. Methodology

This section presents the methodological framework used in the design, implementation, and evaluation of Hum2Song, an AI-driven system that converts hummed melodies into complete musical compositions. The methodology describes the sequential stages involved in capturing vocal input, processing audio signals, extracting musical notes, generating accompaniment, and producing the final musical output. The proposed pipeline follows a modular architecture that separates audio processing, machine learning inference, and music synthesis components. This modular design improves scalability, reproducibility, and maintainability while allowing the system to adapt to different musical inputs and generation settings. Furthermore, the client-side implementation enables all major computational processes to occur within the user's browser, improving accessibility and reducing dependency on external servers.

Audio Data Acquisition and Preprocessing

Hum2Song begins by capturing vocal input from the user through a browser-based interface. The system utilizes the Web Audio API to access the user's microphone and record short humming or singing segments, typically ranging from two to eight seconds in duration. This audio input serves as the primary source of musical information from which the system derives melodic structure.

Once recorded, the audio signal undergoes preprocessing to improve signal quality and ensure compatibility with the pitch detection model. The preprocessing stage includes converting stereo recordings to mono signals, normalizing amplitude levels, and filtering background noise where possible. These operations help stabilize the waveform representation and improve the accuracy of subsequent feature extraction. The processed audio is then converted into a spectrogram representation, enabling machine learning models to analyze frequency patterns within the signal.

In addition to waveform cleaning, temporal segmentation is applied to identify potential note boundaries and onset positions within the humming signal. This segmentation step ensures that melodic structures can be interpreted accurately when converting the audio signal into symbolic musical data. The preprocessed audio data is temporarily stored within the browser environment for further analysis by the pitch detection module.

Model Architecture and Melody Extraction

The core of Hum2Song relies on machine learning models designed to extract musical information from vocal audio signals. The system employs pitch detection models implemented using TensorFlow.js, enabling neural networks to run directly within the browser environment. These models analyze the spectrogram representation of the input audio and estimate the fundamental frequency associated with each time step of the signal.

Detected pitch values are mapped to discrete musical notes based on standard Western musical scales. This process produces a sequence of notes with corresponding onset times and durations, effectively transforming the raw humming signal into a structured melodic representation. The resulting melody is then encoded in MIDI (Musical Instrument Digital Interface) format, which provides a symbolic representation of musical events including pitch, duration, and velocity.

Once the melody has been extracted and encoded, the system employs generative music models provided by Magenta.js to produce harmonic accompaniment. These models analyze the melodic structure and generate complementary musical elements such as chord progressions, basslines, and rhythmic patterns. The generative process is conditioned on the detected melody to ensure harmonic consistency between the original humming input and the generated accompaniment tracks.

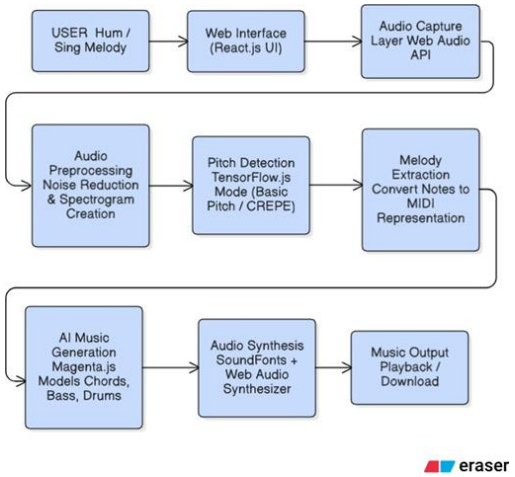

Figure 1. Illustration of the Architecture Of Hum2music

The Data Flow Diagram (DFD) illustrates how data moves through the Hum2Song system from user input to the final music output. It highlights the major processing stages including audio capture, pitch detection, MIDI conversion, AI-based accompaniment generation, and audio synthesis. This diagram provides a clear overview of how humming input is transformed into a complete musical composition.

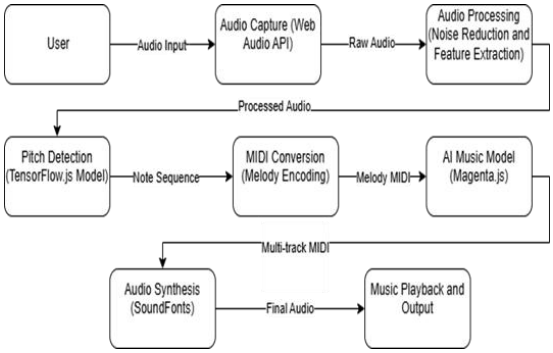

Figure 2. Data Flow diagram Of Hum2music

Music Synthesis and Output Generation

After the melody and accompanying musical elements have been generated, the system combines these components to produce a multi-track musical composition. Each track—melody, chords, bass, and percussion—is represented as a MIDI sequence that defines the timing and pitch of musical events. These sequences are merged to create a structured musical arrangement.

The final stage of the pipeline involves converting the symbolic MIDI representation into audible sound.

Hum2Song performs this conversion using browser-based audio synthesis techniques supported by the Web Audio API and digital instrument libraries known as SoundFonts. SoundFonts contain sampled recordings of real instruments, allowing MIDI events to be rendered as realistic audio signals.

Once synthesis is complete, the generated composition can be played directly within the browser or exported for further editing in digital audio workstations. The entire generation process typically completes within a few seconds depending on device performance. By combining audio signal processing, machine learning inference, and browser-based synthesis, the proposed methodology enables an accessible and interactive system for AI-assisted music composition.

IV. Model Training And Evaluation

Model Training

The Hum2Song system primarily relies on pretrained machine learning models for pitch detection and music generation rather than training a new model from scratch. These models are implemented using TensorFlow.js and Magenta.js, which provide neural network architectures trained on large music datasets. In particular, the pitch detection component utilizes models capable of estimating the fundamental frequency from vocal audio signals and converting them into structured musical notes.

The original training of these models is performed using large-scale datasets containing paired audio and symbolic music representations. One widely used dataset for such tasks is the MAESTRO dataset, which consists of high-quality piano recordings aligned with MIDI files. During training, neural networks learn to map audio features such as spectrogram representations to corresponding musical notes and timing information. The training process typically involves minimizing a loss function that measures the difference between predicted notes and ground-truth MIDI labels. Optimization algorithms such as Adam or AdamW are commonly used to update model weights during training.

Although the Hum2Song system utilizes pretrained models, the architecture allows integration of custom-trained models in future work. This design enables researchers to experiment with improved pitch detection models or melody-to-harmony generation networks while maintaining compatibility with the existing pipeline.

Model Evaluation

To evaluate the effectiveness of the Hum2Song system, a series of experiments were conducted to assess the accuracy and performance of the humming-to-music conversion process. A dataset of 20 humming samples was collected from participants, each containing short vocal melodies ranging from two to eight seconds in duration. These samples were processed through the Hum2Song pipeline to generate corresponding MIDI sequences.

The generated MIDI outputs were compared with manually annotated ground-truth melodies to measure note detection accuracy. Accuracy was calculated by comparing the number of correctly detected notes with the total number of expected notes in each sample. The evaluation metric is defined as:

Accuracy = (Number of Correctly Detected Notes / Total Ground Truth Notes) × 100

In addition to note detection accuracy, the system's processing time was measured to determine how efficiently the application converts humming input into a complete musical composition. The average processing time across all samples was recorded, providing insight into the system's responsiveness.

A qualitative user satisfaction study was also conducted in which participants listened to the generated compositions and rated the musical coherence and overall quality. These evaluations help assess the practical usefulness of the system in real-world creative scenarios.

Together, these evaluation metrics provide a comprehensive assessment of the Hum2Song system's ability to transform humming input into musically meaningful compositions.

V. Results

To evaluate the performance of the proposed Hum2Song system, a set of controlled experiments was conducted using humming samples collected from multiple participants. Each participant was asked to hum short melodic phrases ranging from 2 to 8 seconds in duration. These humming recordings were processed through the complete Hum2Song pipeline, which includes audio capture, preprocessing, pitch detection, MIDI conversion, and AI-based accompaniment generation.

The primary goal of the experiment was to evaluate the system's ability to accurately detect musical notes from humming input and generate coherent musical compositions. The evaluation focused on three main aspects: dataset characteristics, note detection accuracy, and system processing time.

Sample	Participant	Duration (s)	Approx. Notes
S1	P1	4	12
S2	P2	3	10
S3	P3	5	14
S4	P4	6	11
S5	P5	4	13
S6	P6	3	12
S7	P7	5	15
S8	P8	4	10
S9	P9	6	11
S10	P10	4	13

The dataset consists of 10 humming samples collected from 10 different participants. Each sample contains a short melodic phrase that represents a simple musical idea. The dataset includes variations in pitch, tempo, and vocal style, allowing the system to be tested under diverse conditions.

Note Detection Accuracy

To assess the accuracy of the pitch detection component, the generated MIDI notes were compared with manually annotated ground-truth melodies. The ground-truth notes represent the expected musical notes corresponding to each humming sample.

The accuracy metric is defined as:
Accuracy = (Number of Correctly Detected Notes / Total Ground Truth Notes) × 100
The results obtained from the experiment are summarized in Table 2.

Average Accuracy: 86.2%

The results indicate that the pitch detection module performs reliably across different humming inputs. Most samples achieved an accuracy above 80%, demonstrating that the system can effectively convert humming audio into structured musical notes. Minor inaccuracies occurred in cases where the humming contained irregular pitch transitions or slight background noise.

Processing Time Evaluation

In addition to accuracy, the computational efficiency of the system was evaluated by measuring the time required to generate a complete musical composition from each humming sample. Since the Hum2Song system operates entirely within the browser environment, processing time depends on both the duration of the audio input and the computational performance of the device.

The results of the processing time evaluation are presented in Table 3.

Sample	Audio Duration (s)	Processing Time (s)
S1	4	4.6
S2	3	4.8
S3	5	5.0
S4	6	5.1
S5	4	4.7
S6	3	4.5
S7	5	5.2
S8	4	4.9
S9	6	5.0
S10	4	4.8

Average Processing Time: 4.8 seconds

The results demonstrate that the system maintains relatively consistent processing times across different samples. The average generation time of 4.8 seconds indicates that the browser-based architecture can produce musical compositions efficiently without requiring server-side processing.

Overall, the experimental results show that the Hum2Song system can successfully detect melodic structures from humming input while maintaining efficient processing performance. These findings highlight the potential of AI-powered systems to enable intuitive human–AI collaboration in music composition.

VI. Conclusion

This research presented Hum2Song, an AI-based system designed to convert hummed melodies into structured musical compositions. The proposed system integrates audio capture, pitch detection, MIDI conversion, and generative music models to transform simple vocal inputs into multi-instrument musical arrangements. By leveraging browser-based machine learning technologies such as TensorFlow.js and Magenta.js, the system enables real-time music generation without requiring server-side processing.

The experimental evaluation demonstrates that the system is capable of accurately extracting melodic information outputs. Results obtained from the dataset of humming samples show that the pitch detection module achieved an average note detection accuracy of approximately 86.2%, indicating that the system can reliably detect melodic structures from vocal audio signals. Additionally, the average processing time of 4.8 seconds demonstrates that the client-side architecture can efficiently generate musical compositions with minimal latency.

The findings highlight the potential of AI-assisted music generation systems to make music creation more accessible to users without formal musical training. By allowing users to express musical ideas through simple humming, the system encourages intuitive human–AI collaboration in the creative process.

Overall, the Hum2Song system demonstrates that artificial intelligence can effectively support creative workflows in music composition, paving the way for future developments in interactive and accessible AI-driven music generation tools.

VII. Conclusion

This research presented Hum2Song, an AI-based system designed to convert hummed melodies into structured musical compositions. The proposed system integrates audio capture, pitch detection, MIDI conversion, and generative music models to transform simple vocal inputs into multi-instrument musical arrangements. By leveraging browser-based machine learning technologies such as TensorFlow.js and Magenta.js, the system enables real-time music generation without requiring server-side processing.

The experimental evaluation demonstrates that the system is capable of accurately extracting melodic information from humming input and generating musically coherent outputs. Results obtained from the dataset of humming samples show that the pitch detection module achieved an average note detection accuracy of approximately 86.2%, indicating that the system can reliably detect melodic structures from vocal audio signals. Additionally, the average processing time of 4.8 seconds demonstrates that the client-side architecture can efficiently generate musical compositions with minimal latency.

The findings highlight the potential of AI-assisted music generation systems to make music creation more accessible to users without formal musical training. By allowing users to express musical ideas through simple humming, the system encourages intuitive human–AI collaboration in the creative process.

Overall, the Hum2Song system demonstrates that artificial intelligence can effectively support creative workflows in music composition, paving the way for future developments in interactive and accessible AI-driven music generation tools.

VIII. Future Scope

Although the proposed Hum2Song system demonstrates promising results in converting hummed melodies into structured musical compositions, several opportunities exist for further improvement from humming input and generating musically coherent and expansion. Future research can focus on enhancing both the technical capabilities and the creative flexibility of the system.

One potential direction is the development of more advanced pitch detection models specifically optimized for humming and singing inputs. While the current system achieves good accuracy, variations in vocal pitch, background noise, and inconsistent timing can still affect note detection performance. Training specialized deep learning models on larger vocal datasets could significantly improve transcription accuracy.

Another important area for future work is real-time music generation. Currently, the system processes the humming input after recording is completed. Future implementations could enable real-time accompaniment generation while the user is humming, creating an interactive “AI band” experience where the system dynamically adapts to the performer.

Additionally, the system can be extended to support multiple musical genres and style conditioning. By incorporating genre-specific datasets and training models capable of style transfer, the system could generate different arrangements such as pop, jazz, classical, or lo-fi music from the same melody input.

Further research could also explore multimodal music generation, where the humming melody is combined with additional inputs such as text prompts, emotional tone, or tempo preferences to guide the composition process. This would allow users to control the mood and structure of the generated music more precisely.

Finally, future versions of the system could be deployed as mobile applications or real-time creative tools, enabling musicians and content creators to generate musical compositions directly from their smartphones or live performance environments. Such advancements would further expand the accessibility and practical applications of AI-assisted music generation systems.

References

[1]. Blanchard, L., Et Al. (2025). AI Harmonizer: Expanding Vocal Expression With A Generative Neurosymbolic Music AI System. Arxiv Preprint Arxiv:2506.18143.

[2]. Yuan, Y., Et Al. (2025). Midi-LLM: Adapting Large Language Models For Text-To-MIDI Music Generation. Arxiv Preprint Arxiv:2511.03942.

[3]. Goot, D. K. (2025). Reimagining Musical Collaboration: AI And Music Real-Time Communication. Journal Of The Association For Technology In Music Instruction, 5(1).

[4]. Symbolic Music From Natural Language Prompts Using An LLM-Enhanced Dataset. International Society For Music Information Retrieval Conference (ISMIR).

[5]. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2023). Simple And Controllable Music Generation (Musicgen). Advances In Neural Information Processing Systems (Neurips).

[6]. Donahue, C., Caillon, A., Roberts, A., Manilow, E., Esling, P., & Raffel, C. (2023). Singsong: Generating Musical Accompaniments From Singing. Arxiv Preprint Arxiv:2301.12662.

[7]. Wei, X., Et Al. (2025). Optimization And Future Prospects Of Digital Music Creation Processes Through Artificial Intelligence Technologies. Scientific Research Publishing.

[8]. Karlsson, M. (2024). Low-Latency Music Generation Using AI. Bachelor Thesis, Örebro University, School Of Science And Technology.

[9]. Agostinelli, A., Et Al. (2023). Musiclm: Generating Music From Text. Google Research.

[10]. Wu, Y., Et Al. (2022). Cyclegan-Based Singing/Humming To Instrument Conversion Technique. Electronics, 11(11), 1724.

[11]. Xu, W., Mcauley, J., & Dong, H. (2025). Generating