Be the Beat: AI-Powered Boombox for Music Suggestion from Freestyle Dance

Ethan Chang*
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
echang25@mit.edu

Zhixing Chen*
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
zhixingc@mit.edu

Jb Labrune
Media Lab, Design Intelligence Lab
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
labrune@mit.edu

Marcelo Coelho
Department of Architecture, Design Intelligence Lab
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
marceloc@mit.edu

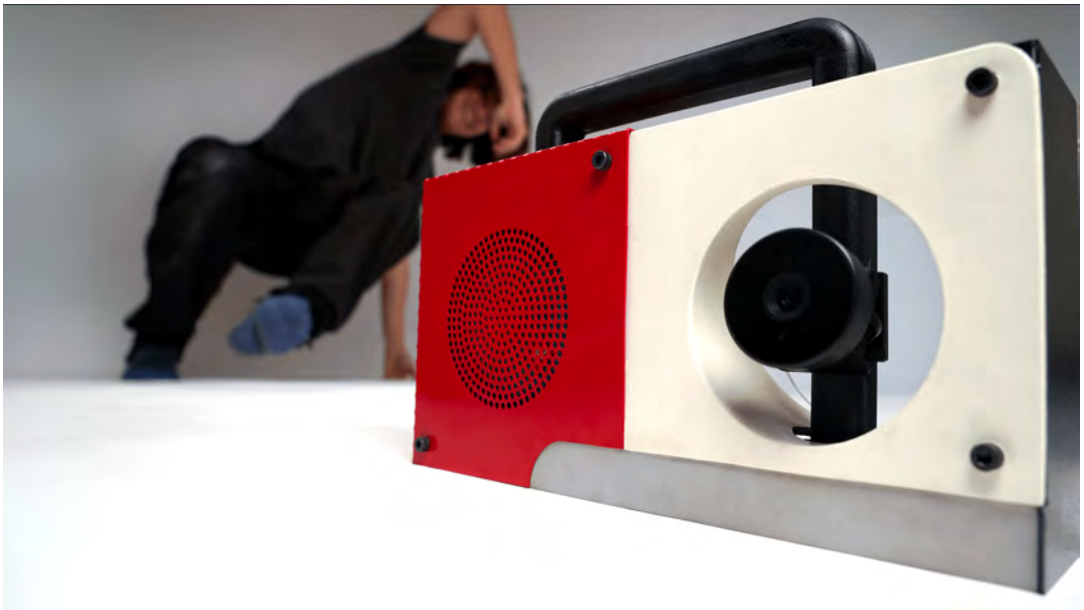


Figure 1: Be the Beat is an interactive boombox that takes dance videos as input through the camera and outputs matching music. The device allows freestyle dancers to guide the quality of music and discover new choreographies.

Abstract

Dance has traditionally been guided by music throughout history and across cultures, yet the concept of dancing to create music is rarely explored. In this paper, we introduce Be the Beat, an AI-powered boombox designed to suggest music from a dancer's movement. Be the Beat uses PoseNet to describe movements for a large language model, enabling it to analyze dance style and query APIs to find music with similar style, energy, and tempo. In our pilot trials, the boombox successfully matched music to the tempo of the dancer's movements and even distinguished the intricacies between house and Hip-Hop moves. Dancers interacting with the boombox reported having more control over artistic expression and described the boombox as a novel approach to discovering dance genres and choreographing creatively. Be the Beat creates a space for human-AI collaboration on freestyle dance, empowering dancers to rethink the traditional dynamic between dance and music.

Keywords

Freestyle Dance, Human-Computer Collaboration, Multimodal AI, Music, Large Language Model

*Both authors contributed equally to this research.

TEI '25, March 04–07, 2025, Bordeaux / Talence, France
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1197-8/25/03
https://doi.org/10.1145/3689050.3705995

ACM Reference Format:
Ethan Chang, Zhixing Chen, Jb Labrune, and Marcelo Coelho. 2025. Be the Beat: AI-Powered Boombox for Music Suggestion from Freestyle Dance. In *Nineteenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '25), March 04–07, 2025, Bordeaux / Talence, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3689050.3705995

1 Introduction

Music and dance are harmonious partners. Typically, music evokes dance, with dancers moving to the rhythm and expressing the melodies through their motions. But what if this dynamic were reversed? Imagine a scenario where subsequent steps in the musical sequence are conditioned on dancers' movements. For example, in Hip-Hop culture, dancers gather in a circle and take turns performing freestyle dances in the middle of the circle. From time to time the music fades out, and this is when the crowd starts making beats either from their voice or body parts to keep the dance circle going based on the dancer's tempo. In breaking battles, the disk jockey (DJ) also creates beats that fit the dancers. These types of interactions between dance movements and music have not been captured by technology before. Be the Beat explores the concept of interactive music-making and integrates AI to replicate the dynamic interplay between a dancer and a beat.

Be the Beat uses a webcam to capture a dance video to pass it through PoseNet, GPT-4, and the Spotify API to suggest music based on the dance movements. Once activated, the device watches the user dancer for 5 seconds and then creates a playlist of suggestions from Spotify based on the user's dance movements. The user can also trigger this recommendation again anytime during the dance, creating an ongoing interaction with the device.

We introduce a novel interaction system that translates dance movements into musical recommendations by leveraging AI systems. Using our AI-powered boombox, the dancers now have control over the type of rhythm, fluidity, and energy of the song that is being suggested. The user can also continuously fine-tune their movements until the song matches what they have in mind. With Be the Beat, new interactions can happen at various scales. Users can also explore new dance genres and choreograph ideas more easily, rather than strictly adhering to a specific song. This device enhances the expressive potential of dance language and empowers dancers with greater control.

2 Related Work

As AI capabilities evolve with more advanced Large Language Models (LLM), Designers have explored various approaches to incorporating LLMs into physical objects [5]. The ways we can implement these models are continually changing. With faster computing and multimodal capabilities, designers can reimagine the use of various inputs and outputs. New design ideas involving the use of music and dance can be achieved without hardcoding specific scenarios through LLMs.

2.1 Physical Music Interfaces

Designers have explored various innovative methods for controlling and generating music. Projects range from basic functions like playing and stopping music [12] to more complex activities such as DJing music [14] and generating new sounds from gestures [9]. In these investigations, the user's motions significantly influence the output from these physical interfaces.

Tomibayashi et al. [14] developed wearable gloves with embedded sensors to allow music control, including fading in and out, mixing, queuing songs, and scratching sound effects. The gestures sensed from the gloves include moving the hands in the x, y, and z axes, clapping, and raising hands. Findings suggest that DJs can accurately control the music in a straightforward way after tuning the sensors, though a request for more functions was also noted. Leonard and Giomi [9] investigated the relationship between gesture and its corresponding sound. An IMU was placed on both the palm and elbow areas to detect motion, intensity, force, and the movement quality of a hand gesture mimicking the strumming of a string.

2.2 Music/Dance Style Classification

Humans can often intuitively identify music genres without substantial training [4], but defining genres in clear, objective terms remains difficult, which limits both human classification and computational models' ability to recognize musical nuances [8]. Music Information Retrieval (MIR) researchers face two major challenges in genre classification. First, the lack of well-defined genre boundaries makes it difficult to establish consistent classifications [8, 11]. Second, current audio-based music classifiers continue to struggle with achieving accuracy rates above 70% , which is the human-level accuracy for classification based on human-defined features [6]. Computers face greater difficulty classifying certain genres than others due to feature overlap and similarities between them [8]. Additionally, music genres are mutable cultural constructs, with definitions evolving, requiring computers to retrain periodically [6].

Newer technology enables the use of dance poses as input data for models, offering the potential to improve genre classification accuracy. Baker et al. [1] extrapolated 10 features, such as joint angular momentum or movement expansion from dance videos, as numeric input to train a machine learning model that maps features to dance genres with up to 76% accuracy without audio input. Mifsud and Inguanez [10] trained a dance style classification model on data from OpenPose, an open-source pose identification package. They used an LSTM Neural Network [13] to feed the dance pose into the machine learning model frame by frame. After training, the model was able to reach 90% accuracy across 10 dance categories.

PoseNet is another widely used method to facilitate dance genre classification. PoseNet is based on a deep convolutional neural network (CNN) that has been trained on a large dataset of images of people in various poses [7]. PoseNet estimates human poses in real-time, detecting key body points from an image or a video stream. It is widely used for applications such as augmented reality, fitness tracking, and motion capture. PoseNet works by analyzing an image and identifying key points on the human body, such as the elbows, knees, shoulders, and hips, to estimate the overall pose.

3 Be the Beat

Be the Beat transforms dance movements into music suggestions through AI technologies, creating a unique, interactive experience. The device can be used when dancers have a specific feeling of dance moves in their mind but can't come up with a song. Triggered by a hand clap from the user, this activates the workflow below:

(1) Dancer wants to find a song that fits their freestyle inspiration in the moment but doesn't have a specific song in mind.
(2) Dancer turns on Be the Beat and claps to start the recording process.
(3) Be the Beat captures live poses in the dancer's performance and processes the data.
(4) Within 5 seconds, the boombox suggests a list of recommended songs from Spotify based on the dance features and starts playing the first song.
(5) Dance hears the song played by the speaker and starts dancing to the selected song.
(6) If the dancer wants another song or shift dance style in certain directions, the dancer can clap and restart the suggestion process to get new recommendations as their freestyle dance develops.

The generative experience of LLMs in music and dance enables users to engage in an ongoing creative dialogue with the machine. While previous physical music interfaces relied on hardcoded mappings between gestures and music synthesis, our project leverages AI capabilities to interpret a broad, open-ended range of gestures and body movements, vastly expanding the input possibilities.

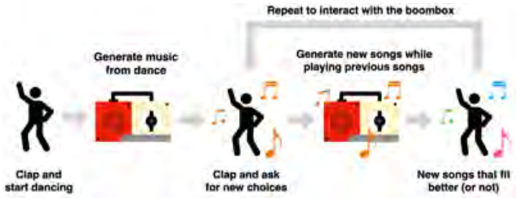

Repeat to interact with the boombox
Generate music from dance
Generate new songs while playing previous songs
Clap and start dancing
Clap and ask for new choices
New songs that fit better (or not)

Figure 2: Storyboard of the interaction cycle. Dancers asking for a new song after songs are suggested, creating a continuous interaction.

3.1 Capturing Pose and Pre-Processing

After the user activates the boombox to record their dances, PoseNet [7] is used to transform human dance gestures into data. PoseNet captures 17 key points on the human body from each frame of the video and stores them in a series of arrays, thereby reducing the amount of information per frame. We record the positions of body key points detected by PoseNet at a framerate of 50 fps for five seconds after activation, with the array storing the x and y coordinates of each key point. To improve the capability of extracting the key features from these arrays, we developed a script to analyze key attributes such as average velocity and moving range. These calculated values are used as input to GPT-4 instead of the full array, enhancing accuracy in identifying dance styles and analyzing BPM.

Table 1: List of metrics used as input to GPT-4 after calculation from the PoseNet data. The data is calculated individually and then gathered as a JSON file. The system processes this input to generate an output of music feature parameters that match the extracted dance metrics.

Metric Name	Definition	Calculated Points
Movement Velocity	Velocity of key body points	All 17 points
Acceleration	Acceleration of key body points	All 17 points
Range of motion	Pixels travel for key body components	Left limb, right limb, left leg, right leg
Joint Angle	Absolute Angle between body parts	Torso and legs, shoulders and back
BPM	Temporal frequency of each key point	All 17 points
Posture distance	Distribution of distance between points	Left hand to right hand, hand to feet, shoulder to ankle, hands to hips
Max distance	Max distance of travel after normalization	All 17 points

Learning from Baker et al. [1], several parameters are chosen as input for GPT-4 as shown in Table 1. For example, we selected movement velocity as a metric to gauge the energy of the dance; the range of motions of different body parts are compared against each other to determine if a particular dance involves more movements in the upper body or lower body. These parameters are also evaluated after normalizing by a bounding box around the dancer to account for differences in shooting angle and distance. The script computes the BPM of each key point using Fourier transform.

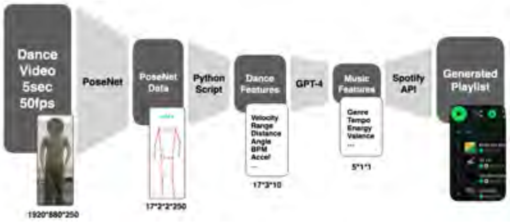


Figure 3: Data processing pipeline of Be the Beat. The pipeline starts by recording the dancer and inputs the video into the system, which is then transformed in steps into PoseNet arrays, dance features, music features, and finally a Spotify playlist.

3.2 GPT-4 Analysis

After transforming the dance video into numerical data, using these data, GPT-4 is utilized to classify the dance style with several reference examples together as input, a technique known as few-shot prompting [2]. The few-shot prompt begins with nine examples of different dance styles from the AIST dance video database [15], which has a variety of dance videos of different styles. The sample videos with labeled dance genres are passed through our preprocessing scripts, each followed by reasoning about the video using a description of the dance, and then the expected output. For example, after prompting GPT-4 with the pre-processed PoseNet data, we add the reasoning: "These data you see is analyzed from a Hip-Hop style dance, known for its upbeat tempo, fast and on beat movements, and bouncy texture in the upper body," and "based on the data you are given, you should output the genre as Hip Hop, with a BPM at 120." Following these example prompts in the prompt line with the provided dance data from the dancer, GPT-4 is asked to infer a likely genre of dance, an energy value from 0-1,

Table 2: List of features that are output from GPT-4 and fed into the Spotify API. Note that some of these features are not tuned in the few-shot prompt. These are also all the possible Spotify API input parameters that are correlated to dance qualities.

Name	Definition
Genre	Dance genre as perceived by GPT4
Tempo	Tempo of the music
Energy	Energy level of the music
Danceability	How danceable is the music
Valence	How positive is the music deemed to be

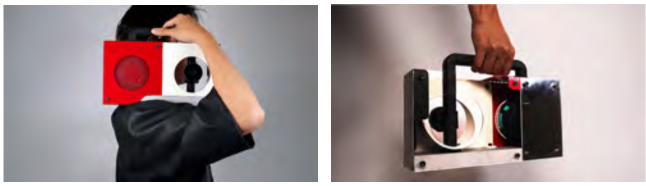
and a likely BPM of the song. These values are then passed onto the Spotify API.

3.3 Spotify Music Recommendation

Spotify is the last step in our pipeline to output music. Using the dance style, energy value, and BPM from the GPT-4 API, the Spotify API can create a playlist that fits the dancer's movement. The parameters that GPT-4 can feed into Spotify are shown in Table 2. The Spotify API then suggests five songs that fit the values GPT-4 provides. Once Spotify suggests a playlist and the dancer starts dancing to the music, they can trigger the boombox again if the music choice is not what they imagined. When the users clap to restart the process, the camera starts recording the dancer again but does not stop the music playing.

3.4 Form Factor

The presence of the boombox is a crucial part of the freestyle dance culture. For the physical appearance, we draw inspiration from the Odd Harmonics theremins by François Chambard [3] due to its attention-drawing color and material choice. The design emphasizes a modern aesthetic, highlighting the connection between the camera, the boombox handle, and the speaker. To highlight the role of the boombox in dance culture, we decided on a multi-material boombox with a bright red to draw attention to the speaker.



(a) Boombox Front View. The model is holding Be the Beat on their shoulder, showing the camera in the right cutout with a handle extending through the body.

(b) Boombox Back View. The clear acrylic back allows the user to see the internals. This shows all the connecting screws and the handle geometry.

Figure 4: Boombox Front and Back Views

4 Evaluation and Discussion

A brief pilot test was conducted involving three participants from our college community. The participants included one female dancer, one male dancer, and one female non-dancer (self-identified), ages 18 to 21. We observed their natural interactions with the device, which they were told was an "AI boombox." We placed Be the Beat in their hands and allowed them to explore its functionality for 10 minutes while they were allowed to observe each other. We instructed users to clap twice to trigger the boombox, which then will suggest music based on their dance movements. Following their interaction with the device, we engaged in a 10-minute conversation to gather their feedback. We categorized the comments and observations into the following categories:

4.1 Ease of use/clarity

Participants initially found the boombox easy to activate and interact with, with one participant stating, "The suggestion function is easy to understand." However, there were moments when a lack of clarity about the system's behavior created confusion. Participants noted that the connection between their movements and the music output was less clear. This resulted in dancers needing to "figure out" how the system translated their movements into music, leading to various attempts.

For example, the non-dancer participant attempted to influence the music tempo by using fast, running-like movements, expecting a faster-tempo song to be suggested, but this was unsuccessful. Moreover, when the second dancer observed the first dancer using the system, they were able to start interacting with it more easily, suggesting that observing others reduced the learning curve. This highlights the need for a demonstration or clearer feedback loops to help users quickly understand the connection between their inputs and the system's outputs.

4.2 Mapping Accuracy

Participants reported mixed experiences with the accuracy of the system's mapping between their movements and the suggested music. For example, a dancer executed specific Hip-Hop moves and received a fitting classic Hip-Hop track. When they switched to house-style footwork, a house song played. Reflecting on this, they remarked, "I think it's challenging for technology to differentiate between house and Hip-Hop, so seeing this distinction was impressive." However, for some trials, the system did not always respond as expected. One dancer commented, "Song choices for me didn't seem to match too well (vibes), other people had better luck." This suggests that while the system was responsive, its algorithm did not consistently meet user expectations for matching genres.

The comments also indicated that some dancers felt their broader intentions (such as energy or mood) were captured, even if the specific song or genre was unexpected. The dancers also appreciated the well-matched tempo. However, the system struggled to differentiate certain nuances, such as in one case where Latin dance was attempted but not successfully queued. These mismatches reveal a need for greater precision in how the system interprets and maps movements to specific musical outputs, especially in terms of genre. While the system shows sufficient mapping ability in certain cases, it does not yet match the users' expectations fully.

4.3 Creativity Enhancement

The unexpected nature of the boombox's outputs was, in some ways, a source of creative enhancement for the dancers. The system's unpredictability pushed participants to adjust their performances, allowing for a playful and exploratory interaction with the system. One dancer remarked, "I didn't expect this song, but it still made sense with my movements," indicating that, although the specific song selection was unexpected, it aligned with their overall creative flow. Participants suggested that the boombox could be useful for discovering new dance styles or experimenting with different choreographic approaches, further supporting its potential as a tool for creative exploration.

However, the other dancer expressed a desire for more control over the interaction. They wanted to be able to guide the system more accurately, particularly in terms of selecting specific genres or tempos to match their intended choreography. The option to tailor music more closely to their artistic goals could enhance both creativity and satisfaction.

4.4 Other Suggested Improvements

Participants suggested several improvements for Be the Beat, including integrating Spotify's API for personalized music recommendations or user playlists, allowing the system to better align with individual preferences. They also proposed enabling users to set preferences based on natural movements or desired genres. Technical enhancements were recommended, such as improving clap detection in larger dance studios where high music volume sometimes caused recognition issues and exploring alternative activation methods like gesture-based commands or video recognition. Furthermore, participants envisioned incorporating facial recognition to gauge emotions, enabling the system to select music that matches the user's mood, highlighting the potential for deeper interaction intelligence.

5 Limitations and Future Work

The unprecedented growth of multimodal AI has accelerated the development of innovative physical interfaces. Through Be the Beat, we created a collaborative dance experience enabled by GPT-4. While the boombox does not yet fully match the users' expectations in music suggestions, it shows partial success in capturing the overall mood and tempo. We would like to improve on:

(1) Model recognition accuracy: Genre classification is a nontrivial task due to overlapping features between genres, evolving definitions of musical styles, and the complexity of accurately distinguishing subtle differences in audio and performance characteristics [6, 8, 11]. Be the Beat is not perfect in this regard, as misclassifications can occur. To address our accuracy, we would fine-tune our model on videos from the AIST database instead of simply few-shot prompting the model. We would also include videos from multiple angles and different lighting situations to ensure the model's accuracy. While we may not achieve 100% accuracy, mismatches can open up new avenues for creativity. Our pilot study showed that unexpected musical pairings often inspired improvisation and new forms of creative expression during performances.

(2) Future evaluation: We recognize that our study, which included only three users and involved brief interactions, is insufficient to fully capture the capabilities of Be the Beat. The small sample size limits the exploration of interactions and experiences with the device. A more comprehensive, long-term investigation with a larger and more diverse group is necessary to understand how AI-powered devices can create or alter dance collaborations. Additionally, we plan to use more in-depth quantitative methods, tracking key metrics such as the success rate of genre classification during each interaction to evaluate the model's recognition accuracy. We will also measure system latency, and record the time spent on each song before switching, which could indicate user preferences. A demonstration on how to use the system will be included to ensure consistent user experience across participants.

6 Conclusion

Be the Beat represents a novel intersection of dance and AI, reversing the traditional dynamic where music dictates movement. By leveraging the capability of models such as PoseNet and GPT-4, we have developed a physical, embedded, and embodied interface that allows dancers to receive music suggestions through their movements, creating an innovative and dynamic experience. Our pilot trials demonstrated that Be the Beat can select music with the tempo and style of various dance genres, providing dancers with greater control over their artistic expression and enabling creative choreography.

We have given the traditional boombox a modern, technological makeover that echoes the concept of creating beats with physical actions. The form factor highlights the connection between the camera and the speaker, while the bold colors add character and bring the artifact to life as an interactive entity. Inspired by past physical music interfaces, Be the Beat opens up possibilities for innovative performance art and offers a tool for both professional dancers and enthusiasts to explore and create in ways previously unimaginable. As we continue to push the boundaries of Be the Beat, we hope to discover transformative applications that will enrich the cultural landscape of dance and music.

Acknowledgments

Thanks to all the designers, dancers, and engineers who helped make this project possible—special thanks to Bill McKenna, Neil Chowdhury and Sarah Schwettmann for valuable inputs and support.

References

[1] Ben Baker, Tony Liu, Jordan Matelsky, Felipe Parodi, Brett Mensh, John W Krakauer, and Konrad Kording. 2024. Computational Kinematics of Dance: Distinguishing Hip Hop Genres. *Frontiers in Robotics and AI* 11 (2024). https://doi.org/10.3389/frobt.2024.1295308

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[3] François Chambard. [n. d.]. Odd Harmonics. Retrieved from https://umproject.com/Odd-Harmonics-theremins.

[4] Elaine Chew, Anja Volk, and Chia-Ying Lee. 2005. Dance Music Classification Using Inner Metric Analysis: A Computational Approach and Case Study Using 101 Latin American Dances and National Anthems. In *The Next Wave in Computing, Optimization, and Decision Technologies*. Springer, Boston, MA, USA, 355–370. https://doi.org/10.1007/0-387-23529-9_23

[5] Marcelo Coelho and Jean-Baptiste Labrune. 2024. Large Language Objects: The Design of Physical AI and Generative Experiences. *interactions* 31, 4 (2024), 43–48. https://doi.org/10.1145/3672534

[6] Mingwen Dong. 2018. Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification. arXiv:1802.09697 [cs.SD] https://arxiv.org/abs/1802.09697

[7] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. *CoRR* abs/1505.07427 (2015). https://doi.org/10.48550/arXiv.1505.07427

[8] D. Kostrzewa, P. Kaminski, and R. Brzeski. 2021. Music Genre Classification: Looking for the Perfect Network. In *Computational Science – ICCS 2021*, M. Paszynski, D. Kranzlmüller, V.V. Krzhizhanovskaya, J.J. Dongarra, and P.M.A. Sloot (Eds.). Lecture Notes in Computer Science, Vol. 12742. Springer, Cham. https://doi.org/10.1007/978-3-030-77961-0_6

[9] James Leonard and Andrea Giomi. 2020. Towards an Interactive Model-Based Sonification of Hand Gesture for Dance Performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Birmingham, UK, 369–374. https://doi.org/10.5281/zenodo.4813422

[10] Yanika Mifsud and Frankie Inguanez. 2021. Dance Style Classification by LSTM RNN. In *IEEE 11th International Conference on Consumer Electronics (ICCE-Berlin)*. Berlin, Germany, 1–6. https://doi.org/10.1109/ICCE-Berlin53567.2021.9720031

[11] Ke Nie. 2022. Inaccurate Prediction or Genre Evolution? Rethinking Genre Classification. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*. Bengaluru, India. https://archives.ismir.net/ismir2022/paper/000039.pdf

[12] Josef Roth, Jan Ehlers, Christopher Getschmann, and Florian Echtler. 2021. TempoWatch: a Wearable Music Control Interface for Dance Instructors. In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '21)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3430524.3442461

[13] Ralf C. Staudemeyer and Eric Rothstein Morris. 2019. Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks. arXiv:1909.09586 [cs.NE] https://arxiv.org/abs/1909.09586

[14] Yutaka Tomibayashi, Yoshinari Takegawa, Tsutomu Terada, and Masahiko Tsukamoto. 2009. Wearable DJ system: a new motion-controlled DJ system. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology (ACE '09)*. Association for Computing Machinery, New York, NY, USA, 132–139. https://doi.org/10.1145/1690388.1690411

[15] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. 2019. AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*. Delft, Netherlands. https://doi.org/10.5281/zenodo.3532606