HEARTBEAT-ADAPTIVE MUSIC RECOMMENDATION

Designing a Physiologically Responsive Music Player on the MTG-Jamendo Corpus

Oğuzhan Tuğral

otugral@umass.edu

Abstract

This study presents a user-experience analysis of a heartbeat-adaptive music recommendation system built on the MTG-Jamendo Dataset. The system performs real-time music selection from a catalogue of 18,486 tracks using heart rate (HR) and heart rate variability (HRV) as its only input signals. In this framework, cardiovascular dynamics constitute a biologically grounded interaction modality. The core challenge is a mapping problem: translating continuous physiological signals into discrete and musically meaningful selections within an affectively structured corpus. Addressing this requires both reliable signal interpretation and a consistent representation of musical affect. A central contribution lies in extending the dataset itself. The original MTG-Jamendo corpus is augmented through the systematic addition of affective and computational tags, transforming it into a dynamically queryable and physiologically responsive structure. Thus, the work integrates corpus design and system implementation. The study proceeds by identifying user pain points, reviewing dataset limitations, introducing a valence–energy (V, E) annotation layer, and formalizing a physiological mapping model. A deterministic lookup table maps 59 mood categories onto Russell's Circumplex Model of Affect, while the tagging schema ensures consistent retrieval. Consequently, the system enables interpretable, reproducible, and real-time music selection grounded in physiological state.

Contents

1 Introduction

Most music recommendation systems are driven by *historical behaviour*: what the listener played before, what listeners with similar profiles played, or what is currently popular. These approaches optimise for engagement and familiarity, but they have no mechanism for responding to *what the listener's body is doing right now*. A person whose heart rate is elevated after exercise receives the same recommendations as the same person at rest. A person in a state of autonomic stress receives the same queue as a person in a parasympathetically dominant, restful state.

The application presented in this work is built around a different premise: that the most relevant signal for music selection at any given moment is not listening history but *current physiological state*. The listener's HR and HRV are measured continuously; these values determine which mood zone the listener currently occupies; and the system selects tracks from the MTG-Jamendo corpus whose affective coordinates match that zone. The loop closes when the music shifts the listener toward a *target physiological state* — relaxation after exertion, gentle arousal from fatigue, stabilisation from stress.

Section 2 documents the four user pain points that motivated this design. Section 3 describes the MTG-Jamendo corpus as originally published and identifies the gap that required the authors' annotation contribution. Section 4 specifies that contribution: the V/E annotation layer and its design rationale. Section 5 presents the physiological mapping model in full, from raw HR and HRV to track selection. Section 6 articulates the UX design principles that emerged from the project.

2 User Pain Points

Four pain points were identified through a review of existing music recommendation systems, wearable health application patterns, and the music–cognition literature.

2.1 Pain Point 1 — Music Recommendations Are Temporally Disconnected from Physiological State

Existing recommender systems — whether based on collaborative filtering, content- based features, or large language model embeddings — have no access to the listener's real-time physiological state. A runner at 170 bpm who finishes a session and wants to cool down cannot tell Spotify "match my heartbeat now"; they can only skip tracks manually until something fits. A person with chronic anxiety who wants music calibrated to their current HRV level has no such option at all. The recommender does not know the body; it only knows the history.

Design response: Make the wearable sensor the primary input device. HR and HRV are sampled continuously; every selection decision is made against the listener's current cardiovascular state, not their stored profile.

2.2 Pain Point 2 — Mood-Based Filters Require Active Cognitive Effort

Many streaming platforms offer mood-based playlists ("Chill", "Focus", "Energy Boost"). But selecting a mood requires the listener to *self-assess* and *manually navigate* — two acts of cognitive effort that are especially demanding precisely when the listener is in a state where self-regulation is most needed (post-exertion, under stress, trying to fall asleep). The interface assumes a reflective, deliberate user; the physiological user is often in none of those states.

Design response: Eliminate the mood-selection step entirely. The cardiovascular signal performs the self-assessment automatically. The listener does not choose a mood; their

body reports one, and the system acts on it.

2.3 Pain Point 3 — No Principled Bridge between Affective Music Tags and Physiological Parameters

Even for researchers or developers who wish to build physiologically responsive music systems, no standard, interpretable mapping exists from mood or affective tags to cardiovascular parameters. Spotify's audio features (valence, energy, danceability) are computed by undisclosed machine learning models and are not designed to be interpreted in physiological terms. Tags on open platforms such as Jamendo are categorical and uncoordinated. There is no published, auditable function of the form: *"a listener whose HR is X and HRV is Y should receive music tagged Z".*

Design response: Construct and publish such a function explicitly. The V/E annotation layer (Section 4) and the mapping model (Section 5) together constitute this function, expressed in terms that a physiologist, a music therapist, or a software engineer can inspect and modify.

2.4 Pain Point 4 — Existing Affective Music Corpora Are Not Usable as Real-Time Selection Pools

Academic datasets for music emotion recognition (DEAM, PMEMO, MoodSwings) contain continuous valence–arousal annotations, but their catalogues are small (hundreds of tracks), their audio is often not freely distributable, and they were designed for model training rather than real-time playback. The MTG-Jamendo corpus is large (18,486 tracks), openly licensed, and structurally rich; but in its original form it provides only categorical mood/theme tags with no quantitative affective coordinates, making it unusable as a direct selection pool for a physiologically parameterised recommender.

Design response: Extend the MTG-Jamendo corpus with a quantitative V/E annotation layer (Section 4), transforming it into a real- time-selectable affective catalogue.

3 The MTG-Jamendo Corpus: Original Data Structure

3.1 Provenance and Scale

The MTG-Jamendo Dataset (Bogdanov, Porter, Tovstogan, & Won, 2019) is an open dataset developed at the Music Technology Group, Universitat Pompeu Fabra, and released for the Emotion and Theme Recognition in Music Task of the MediaEval 2019 benchmark. It contains *18,486 audio tracks* drawn from the Jamendo platform, which hosts Creative Commons-licensed music from independent artists. The corpus spans *93 genre categories* and is the largest freely distributable affective music dataset in terms of full-length track count.

3.2 Original Annotation Structure

In its published form, each record in the MTG-Jamendo corpus carries:

- a *Jamendo track ID* (`jid`): a unique integer identifier
- a *mood/theme tag list* (`moods`): one or more categorical string labels drawn from a vocabulary of **56 mood/theme categories**, crowdsourced from social media tags on the Jamendo platform
- a *genre tag list* (`genres`): one or more categorical genre labels drawn from a vocabulary of 87 genre categories

The annotation is *categorical and tag-based*. The mood tags — such as `happy`, `dark`, `epic`, `relaxing` — represent listener-assigned or artist-assigned semantic labels. They are not grounded in any psychometric model and carry no quantitative affective coordinates. The dataset does not include valence or arousal scores. Table 1 summarises the original schema.

Table 1: MTG-Jamendo corpus: original published schema (Bogdanov et al., 2019)

Field	Type	Content in original corpus
`jid`	integer	Unique Jamendo track identifier
`moods`	string[]	0–n categorical mood/theme tags (crowdsourced)
`genres`	string[]	0–n categorical genre tags (crowdsourced)
`v`	*absent*	*Not present in original dataset*
`e`	*absent*	*Not present in original dataset*

3.3 The Gap: No Quantitative Affective Coordinates

Because the MTG-Jamendo corpus provides only categorical tags, it cannot be used directly as the selection pool for a physiologically parameterised recommender. Cardiovascular signals are *continuous*; categorical tags are *nominal*. There is no natural distance function between a tag like `dark` and a HR value of 85 bpm. To make the corpus selectable by physiological state, a quantitative affective layer is required. This is the gap that the authors' contribution addresses.

4 Authors' Contribution: The Valence–Energy Annotation Layer

4.1 Design Rationale

The authors extended the MTG-Jamendo corpus by assigning each of the *59 mood categories* present in the working subset a coordinate pair (V, E) in *Russell's two-dimensional circumplex model of affect* (Russell, 1980), where:

- $V \in [0, 1]$: *valence* — the positive–negative affective dimension (0 = maximally unpleasant; 1 = maximally pleasant)
- $E \in [0, 1]$: *energy/arousal* — the activated–deactivated dimension (0 = maximally calm; 1 = maximally aroused)

The assignment was made *by hand*, using the semantic content of each mood label

as interpreted against the circumplex literature, validated against published lexicon-to-circumplex mappings (e.g., Warriner et al., 2013; Vo et al., 2009) and cross-checked for internal consistency across the 59 entries.

4.2 Key Structural Properties of the Annotation Layer

The resulting annotation has three properties that are consequential for the application's design:

Full determinism. Each of the 59 mood categories maps to *exactly one* (V, E) pair. No mood label is associated with more than one coordinate. This determinism is by design: it makes the system *reproducible* (two instances of the application running the same mood label will always obtain the same coordinates), *auditable* (the full mapping is inspectable as a 59-row table), and *correctable* (a researcher who disagrees with the placement of `dramatic` in the circumplex can update a single table entry).

Quantised values. Valence values are drawn from 27 discrete levels in [0.10, 0.85]; energy values from 29 discrete levels in [0.15, 0.90]. This quantisation reflects the resolution at which mood labels can be meaningfully differentiated in the circumplex and prevents false precision.

Corpus-level coverage. Because every track in the working subset carries exactly one mood tag, and every mood tag has a (V, E) coordinate, every track in the corpus now has a quantitative affective position. The corpus is fully covered; there are no unresolvable records.

4.3 The Extended Schema

The authors' contribution transforms each corpus record into the extended schema shown in Table 2.

Table 2: Extended corpus schema after authors' V/E annotation layer

Field	Type	Content
`jid`	integer	Unique Jamendo track identifier
`moods`	string[]	Categorical mood/theme tag (original)
`genres`	string[]	Categorical genre tag (original)
`v`	float	**[Authors' addition]** Valence coordinate; 27 discrete levels in [0.10, 0.85]
`e`	float	**[Authors' addition]** Energy/arousal coordinate; 29 discrete levels in [0.15, 0.90]

4.4 The Full Mood-to-(V, E) Lookup Table

Table 3 presents the complete 59-entry lookup table, sorted by valence from lowest (most negative affect) to highest (most positive affect). The energy axis varies independently within each valence region, capturing the important empirical fact — confirmed by the circumplex literature — that negative valence does not imply low arousal: `horror` (V=0.10, E=0.65) and `heavy` (V=0.18, E=0.80) sit in the low-valence, high-energy quadrant, while `sad` (V=0.18, E=0.22) and `melancholic` (V=0.22, E=0.25) sit in the low-valence, low-energy quadrant.

Table 3: Complete mood-to-(V, E) lookup table; 59 entries; sorted by V ascending

Mood label	V	E	Track count
horror	0.10	0.65	87
sad	0.18	0.22	338
heavy	0.18	0.80	109
dark	0.20	0.55	1,142
melancholic	0.22	0.25	212
dramatic	0.25	0.62	184
drama	0.28	0.50	263
ballad	0.35	0.28	326
emotional	0.38	0.42	817
slow	0.40	0.20	149

Table 3 continued

Mood label	V	E	Track count
powerful	0.40	0.82	58
deep	0.42	0.35	603
trailer	0.42	0.80	39
space	0.45	0.20	238
movie	0.45	0.50	155
film	0.45	0.55	869
epic	0.45	0.78	527
meditative	0.48	0.15	442
soundscape	0.48	0.20	315
documentary	0.48	0.45	300
ambiental	0.50	0.18	458
background	0.50	0.28	468
game	0.50	0.72	164
action	0.50	0.85	407
calm	0.52	0.18	418
relaxing	0.55	0.20	609
nature	0.55	0.22	94
soft	0.55	0.25	55
melodic	0.55	0.45	809
adventure	0.55	0.75	375
fast	0.55	0.90	72
dream	0.58	0.25	770
mellow	0.58	0.30	60
retro	0.58	0.55	125
cool	0.60	0.55	229
corporate	0.60	0.55	276

Table 3 continued

Mood label	V	E	Track count
sport	0.60	0.82	77
sexy	0.62	0.45	49
travel	0.62	0.52	47
commercial	0.62	0.58	264
romantic	0.65	0.35	191
advertising	0.65	0.62	639
energetic	0.65	0.88	1,075
holiday	0.68	0.48	21
motivational	0.68	0.70	60
groovy	0.68	0.78	141
christmas	0.70	0.50	520
hopeful	0.70	0.55	84
party	0.70	0.90	141
love	0.72	0.40	540
inspiring	0.72	0.60	190
upbeat	0.72	0.82	91
positive	0.75	0.65	80
summer	0.75	0.65	78
uplifting	0.75	0.68	88
fun	0.78	0.75	312
funny	0.80	0.70	100
children	0.82	0.65	398
happy	0.85	0.72	738

5 Physiological Mapping Model: HR and HRV to Track Selection

5.1 Overview of the Mapping Problem

The core engineering problem of the application is to define a function:

$$f\big(\mathrm{HR}(t),\ \mathrm{HRV}(t)\big) \longrightarrow \text{track} \in \mathcal{C} \tag{1}$$

where $\mathrm{HR}(t)$ is the listener's heart rate at time t (beats per minute), $\mathrm{HRV}(t)$ is a heart rate variability metric at time t, and $\mathcal{C}$ is the extended MTG-Jamendo corpus. The function must be *continuous* (small changes in HR or HRV should produce smooth changes in music selection), *interpretable* (a musician or therapist should be able to understand and predict its behaviour), and *real-time* (it must execute faster than the sensor sampling interval).

The mapping proceeds in three stages: (1) cardiovascular state estimation, (2) target affective coordinate computation, and (3) corpus query and track selection.

5.2 Stage 1 — Cardiovascular State Estimation

Heart rate (HR). HR is measured as instantaneous beats per minute from the wearable sensor. It is normalised against the listener's personal resting baseline $\overline{\mathrm{HR}}_{\mathrm{rest}}$, established during a 60-second calibration period at session start:

$$\Delta\mathrm{HR}(t) = \mathrm{HR}(t) - \overline{\mathrm{HR}}_{\mathrm{rest}} \tag{2}$$

The deviation $\Delta\mathrm{HR}$ is the primary driver of the energy axis target (Section 5.3). Using a deviation rather than an absolute value accounts for inter-individual differences in resting HR.

Heart rate variability: the H9 metric. HRV is operationalised using the *H9 metric*, defined as the proportion of successive RR intervals (inter-beat intervals, in milliseconds) that differ from the immediately preceding interval by more than 9 ms:

$$\mathrm{H9}(t) = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathbf{1}\big[\, |RR_i - RR_{i-1}| > 9 \text{ ms} \,\big] \tag{3}$$

where N is the number of RR intervals in a 30-second sliding window. H9 is preferred over the standard RMSSD metric for three reasons. First, H9 is *robust to short windows*: as a proportion-based statistic it remains interpretable with as few as 8–12 beats, whereas RMSSD is sensitive to single outlier intervals at short window lengths. Second, H9 has a natural *probabilistic interpretation*: it is the empirical probability that the next heartbeat will deviate non-trivially from the preceding one, mapping intuitively onto the concept of cardiac irregularity. Third, H9 is *computationally lightweight*: it requires only comparison and counting, with no exponentiation, making it suitable for low-latency embedded sensor platforms.

H9 is normalised against the listener's calibration baseline $\overline{\mathrm{H9}}_{\mathrm{rest}}$:

$$\Delta\mathrm{H9}(t) = \mathrm{H9}(t) - \overline{\mathrm{H9}}_{\mathrm{rest}} \tag{4}$$

Positive ΔH9 indicates *parasympathetic dominance* relative to baseline (higher beat-to-beat variability, lower arousal); negative ΔH9 indicates *sympathetic activation* (more regular, clock-like heartbeat, higher arousal or stress).

5.3 Stage 2 — Target Affective Coordinate Computation

The two normalised cardiovascular signals are mapped to target affective coordinates $(V_{\mathrm{target}},\ E_{\mathrm{target}})$ in the circumplex.

Energy target from HR. The energy axis target is derived primarily from ΔHR:

$$E_{\text{target}} = 0.50 + \alpha \cdot \Delta\text{HR} \tag{5}$$

where $\alpha = 0.015$ (per bpm; calibrated so that a +20 bpm deviation maps to $E_{\text{target}} = 0.80$, placing the listener in the high-energy selection zone). The target is clamped to $[0.15,\ 0.90]$ to stay within the corpus range.

The intuition is direct: a high HR indicates sympathetic activation, and the application should select music whose energy matches or — depending on the application mode — gently counters that activation level.

Valence target from H9. The valence axis target is derived from ΔH9:

$$V_{\text{target}} = 0.55 + \beta \cdot \Delta\text{H9} \tag{6}$$

where $\beta = 1.20$ (calibrated so that a ΔH9 of +0.20, a strongly parasympathetic state, maps to $V_{\text{target}} \approx 0.79$, in the high-valence region). The target is clamped to $[0.10,\ 0.85]$.

The rationale is grounded in the autonomic–emotion literature: high HRV is associated with positive affect and emotional regulation capacity; suppressed HRV is associated with negative affect and stress. The valence target therefore tracks the listener's inferred affective tone rather than their arousal level.

Application modes. The mapping can be run in two modes that alter the relationship between physiological state and selection target:

- *Mirror mode:* select music whose $(V,\ E)$ coordinates are closest to $(V_{\text{target}},\ E_{\text{target}})$. The music matches the listener's current state.
- *Guide mode:* select music displaced toward a predefined *goal state* $(V_{\text{goal}},\ E_{\text{goal}})$ (e.g., relaxation: $V = 0.55$, $E = 0.20$). The music gently leads the listener from their current state toward the goal.

Guide mode is the primary clinical use case (post-exercise recovery, stress reduction, sleep preparation); Mirror mode is the primary exploratory or research use case.

5.4 Stage 3 — Corpus Query and Track Selection

Given $(V_{\text{target}},\ E_{\text{target}})$, the system queries the extended corpus for the track(s) minimising the weighted Euclidean distance in the circumplex:

$$d(V_{\text{target}},\ E_{\text{target}},\ \text{track}_k) = \sqrt{w_V \cdot (V_{\text{target}} - V_k)^2 + w_E \cdot (E_{\text{target}} - E_k)^2} \tag{7}$$

where V_k and E_k are the corpus coordinates of track k, and $w_V = 1.0$, $w_E = 1.5$ by default (energy is weighted more heavily because it is the more physiologically proximate variable). The track with minimum d is queued for playback. A minimum play duration of 30 seconds is enforced before re-querying, preventing rapid switching during transient HR fluctuations.

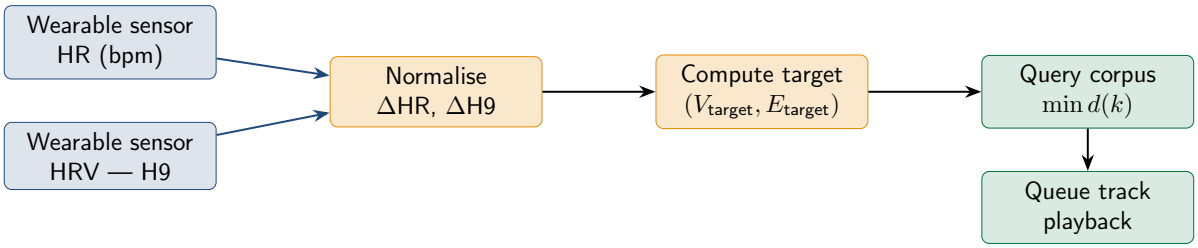

Figure 1: Three-stage physiological mapping pipeline. HR and HRV are the sole inputs; the extended MTG-Jamendo corpus is the sole selection pool.

5.5 Cardiovascular Zones and Corpus Coverage

To provide an interpretable overview of the corpus coverage, Table 4 groups the 59 mood categories into four *cardiovascular zones* defined by energy (E), which is the axis most directly driven by HR. Within each zone, valence (driven by H9) determines which sub-region of moods is selected.

Table 4: Cardiovascular zones by energy level, with mood categories and track counts

Zone	E range	Tracks	Mood categories
Resting	$E < 0.30$	4,892 (26%)	meditative, calm, ambiental, relaxing, space, soundscape, slow, sad, nature, melancholic, soft, dream, background, ballad
Light	$0.30 \leq E < 0.55$	4,375 (24%)	mellow, deep, romantic, love, emotional, melodic, documentary, sexy, holiday, drama, christmas, movie, travel
Moderate	$0.55 \leq E < 0.75$	5,795 (31%)	film, cool, dark, retro, hopeful, corporate, commercial, inspiring, advertising, dramatic, children, horror, positive, summer, uplifting, funny, motivational, happy, game
High	$E \geq 0.75$	3,424 (19%)	adventure, fun, epic, groovy, heavy, trailer, sport, upbeat, powerful, action, energetic, fast, party

The corpus is densest in the Moderate zone (31% of tracks), reflecting the genre composition of the Jamendo platform (ambient, classical, electronic). The Resting zone is the second largest (26%), providing ample material for relaxation and sleep-preparation use cases.

A cross-cutting *stress quadrant* — low valence combined with high energy ($V < 0.45$, $E > 0.60$) — contains five mood categories: `horror` (V=0.10, E=0.65), `heavy` (V=0.18, E=0.80), `dramatic` (V=0.25, E=0.62), `trailer` (V=0.42, E=0.80), and `powerful` (V=0.40, E=0.82). These moods aggregate 477 tracks and correspond to the physiological profile of acute stress: elevated HR, suppressed HRV. In Guide mode, the application will steer *away* from this quadrant toward higher valence at equivalent or lower energy.

6 UX Design Principles

The development of this application surfaces four principles for physiologically responsive UX design.

1. The body is the interface. The application has no mood selector, no skip button in the traditional sense, and no playlist. The listener's cardiovascular state performs all navigation. This eliminates a class of interaction — "what should I listen to now?" — that is most cognitively demanding precisely when the listener is least capable of deliberate self-reflection (post-exertion, under stress).

2. Interpretability is a first-class requirement in health-adjacent UX. Because the application's selections affect the listener's physiological state (music is a non-pharmacological intervention), a black-box model is unsuitable. Every selection decision must be traceable to a specific HR value, a specific H9 value, and a specific lookup table entry. The V/E annotation layer and the mapping equations are the interpretability mechanism; they are not an implementation detail.

3. Calibration personalises without profiling. The 60-second resting calibration at session start establishes personal baseline HR and H9, ensuring that the mapping responds to *deviations* from the individual's own physiology rather than to population norms. No persistent listener profile is stored; calibration is ephemeral and session-local.

4. A fixed lookup table is a design asset for reproducibility. The deterministic mood-to-(V, E) mapping may appear to be a limitation compared to a learned affective model. In a health-adjacent application it is an advantage: two listeners in identical physiological states will receive music from the same region of the corpus. Results are reproducible across sessions and across users, enabling future clinical or research validation.

7 Conclusion

This work has presented the UX research foundation of a heartbeat-adaptive music recommendation application. Four user pain points — temporal disconnection from physiological state, cognitive overhead of mood-based navigation, the absence of a principled mood-to-physiology bridge, and the unusability of existing affective corpora for real-time selection — motivated a design in which the listener's HR and HRV are the sole interface.

The application's data foundation is the MTG-Jamendo corpus, extended in the present work: a 59-entry valence–energy lookup table that assigns every mood category in the corpus a unique coordinate in Russell's circumplex. This annotation layer is fully deterministic, yet it remains extensible through future data-driven refinement based on user feedback. Moreover, it is transparent and reproducible—properties that are not merely desirable but fundamentally required in the health-adjacent context in which the application is intended to operate.

The physiological mapping model translates HR deviation (ΔHR) to an energy axis target via a linear function (Eq. 5), and HRV deviation (ΔH9) to a valence axis target via a second linear function (Eq. 6). A weighted Euclidean distance query over the extended corpus (Eq. 7) selects the track whose affective position most closely matches the target. In Guide mode, the target is displaced toward a predefined goal state, enabling the application to function as a gentle physiological regulation tool.

The result is a system in which the design question "what music should play now?" is answered entirely by the listener's body.