Progress Report: A Music Recommendation System for Emotional Regulation via Reinforcement Learning

1. RL System Design Methodology

1.1 Emotional and Musical Framework

To allow the system's operation, it is crucial to create a framework transforming human emotions into a format readable by the RL agent.

The theoretical foundation here is Russell's Circumplex Model of Affect [2], which depicts emotions in a two-dimensional space of Valence (pleasure/displeasure) and Arousal (activation).

A set of 12 emotions (e.g., "Happy," "Sad," "Calm," "Tense") has been created for user engagement. Building on the TFG [1], this choice provides more detail by addressing all four quadrants of Russell's model.

Empirical V-A Values: The V-A values for each emotion label are derived from empirical affective standards from psycholinguistics, notably the work of Warriner et al. (2013) [4], hence providing the system more scientific rigour in a significant improvement over the TFG. For use in the model, the values are normalized to the range [-1, 1].

Following a review of music psychology research [5, 6, 7, 8, 9], five musical characteristics were chosen for the RL agent to directly regulate: valence, energy (arousal), tempo, mode, and danceability. Other characteristics (such as acousticness or instrumentalness) have been temporarily excluded from immediate agent control, delegating them to possible heuristics in the API query layer to keep a manageable action space.

1.2 Reinforcement Learning Problem Formulation (MDP)

Formalizing the issue as a Markov Decision Process (MDP) is:

- State Space (S): An 9-dimensional continuous vector presenting the agent a full perspective of the present circumstances:
    1. The user's V-A coordinates now are
    2. Vtarget and Atarget: The session's emotional objective.
    3. HA and HV: An exponentially weighted moving average of the V-A history gives a feel of the past direction.
    4. Last interaction information: Play ratio, a skip indication, and an explicit feedback indicator (like/dislike).

- Action Space (A): A 5-dimensional continuous vector, Action Space (A) is the one where the agent establishes the perfect musical profile for the upcoming song. Every element is normalized:
  1. Target valence for the song (-1, 1).
  2. Target the arousal for the song ([-1, 1]).
  3. Aimed song tempo: [0, 1].
  4. For the music's target mode, [0, 1] is interpreted as Minor/Major.
  5. Aim danceability for the song ([0, 1]).
- The agent was meant to be effectively guided combining several goals with varying weights:

$$Rt = wtrans \cdot Rtrans, t + weng \cdot Reng, t + wfeed \cdot Rfeed, t + Rstep$$

  where the parts reward progress toward the V-A target (Rtrans), good participation (Reng), and positive explicit user feedback (Rfeed), while a little penalty per step (Rstep) promotes efficiency.

1.3 Selected Algorithm (PPO)

The Proximal Policy Optimization (PPO) approach was selected. It is a strong and well-known Actor-Critic approach. With its great equilibrium of performance, stability, and simplicity of tuning, it is extremely well suited for ongoing action environments such the one specified for this project.

2. Implementation and Training

2.1 Simulated Environment (*UserEmotionalSimulator.py*)

Given the impossibility of training an RL agent from scratch with millions of real user interactions, the first major implementation step was the creation of a custom simulation environment using the Gymnasium library.

- **Purpose:** To provide a fast, safe, and controllable testbed for the agent to learn the abstract policy of emotional navigation.

- **Internal Logic:** The simulator models the dynamics of a "virtual user." It receives the musical profile proposed by the agent and, based on configurable parameters (like "emotional susceptibility" and "noise"), it calculates the user's new V-A position. It also probabilistically simulates user interactions like skipping a song or providing explicit feedback, enabling the calculation of the composite reward.

- **Validation:** The environment was debugged and validated using Stable Baselines3's check_env utility to ensure its compatibility with the PPO algorithm.

2.2 PPO Agent Training Process

Once the simulator had been verified, agent training started.

For the implementation of the PPO agent, the Stable Baselines3 library was utilized.

Common initial hyperparameters for PPO were established (learning rate, gamma, etc.), along with a neural network architecture (MlpPolicy).

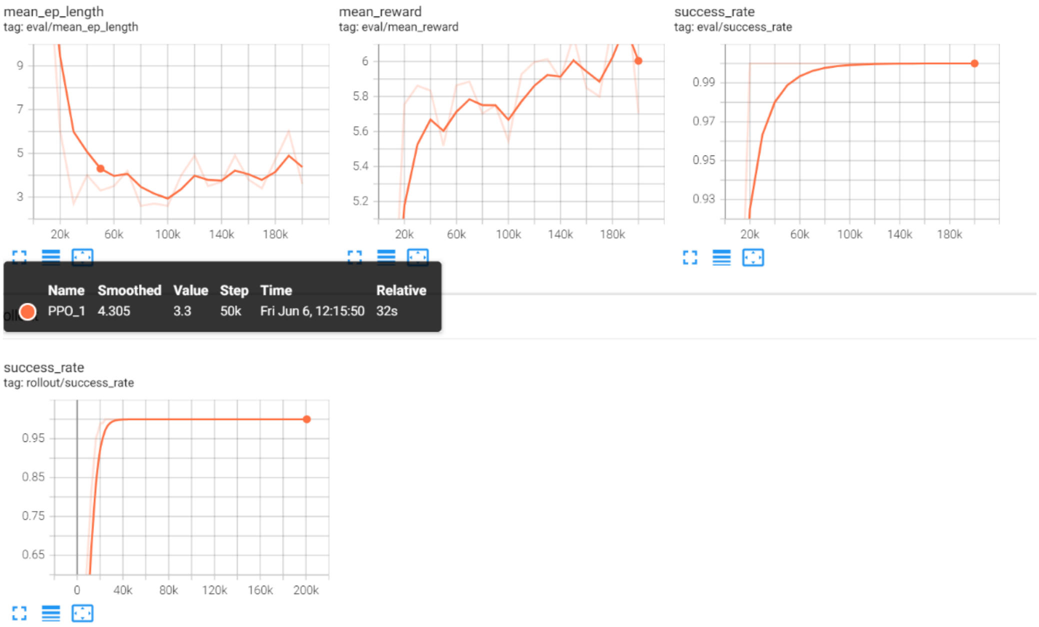
Monitoring: TensorBoard was used to watch the training process, recording important measures including mean reward per episode, episode length, and success rate.

3. Preliminary Results and Analysis

The training of the PPO agent in the simulated environment has yielded highly positive and promising results.

3.1 Analysis of Learning Curves

The agent's training was monitored using TensorBoard over 200,000 timesteps. The key metrics, particularly from the periodic evaluations (eval/), demonstrate successful and efficient learning.



	Name	Smoothed	Value	Step	Time	Relative
●	PPO_1	4.305	3.3	50k	Fri Jun 6, 12:15:50	32s

- The most important performance indicator is mean reward (*eval/mean_reward*). With the average reward per episode rising from around 5.2 to over 6.0, the graph reveals a definite and consistent upward trend. This upward slope validates the agent's successful learning of policy optimizing by performing acts that maximize the total reward function. In brief, the agent is really learning to complete the given assignment.

- **Success Rate** (*eval/success_rate*): This indicator shows the proportion of evaluation episodes during which the agent efficiently directs the virtual user to the intended emotional condition. The curve indicates very fast learning, reaching a success rate of over 95% before 40,000 timesteps and stabilizing around 99%. This shows that by fulfilling the main goal, the agent promptly determines the main plan to get the big end-of-episode reward (R_target_reached_bonus).
- **Mean Episode Length** (*eval/mean_ep_length*): This graph shows a more advanced learning approach. It shows two different phases:
    1. Efficiency Optimization Phase (0k - 80k steps): From about 9 steps to a low of about 3.5, the average episode length decreases dramatically. The agent learns the most straight path to the goal during this phase so as to guarantee the last bonus and avoid amassing step penalties.
    2. Reward Refinement Stage (80k–200k steps): Once efficiency is mastered, the episode length grows somewhat and levels around 4–5 steps. This conduct indicates that the agent, having mastered how to consistently reach the objective, starts to look for methods to maximize the whole episode reward. It probably finds that moving a few more steps with musical profiles that evoke positive intermediate rewards (e.g., from simulated "likes" or non-skips) produces a final score higher than just sprinting to the finish line. This points to a more subtle and smart learned policy.

3.2 Qualitative Analysis of the Learned Policy

Loading the best-trained model into a test script, a qualitative analysis of the agent's acquired strategy was carried out. The goal was to investigate the sequence of abstract musical profiles (that is, the agent's action vectors) produced for several emotional changes. The results show that the agent has developed consistent and naturally correct methods:

- From "Sad" to "Happy," the agent consistently produces a series of profiles with ever higher Arousal and Valence, therefore showing a rational "uplifting" strategy.
- From "Tense" to "Calm": The agent learns to present profiles with lowering arousal and slowly rising valence, which fits a reasonable relaxation approach.
- From “Happy” to “Content”: The agent very well grasps the complexity of this change, suggesting profiles that retain high Valence while lowering Arousal, so reaching the objective effectively in very few steps.

This study confirms that the agent's developed method of negotiating the abstract V-A space is coherent and efficient. A step now beset with technological problems related to API query formation, the next crucial phase of the project is translating these produced musical profiles into actual song suggestions using the Spotify API and verifying their real-world relevance.

4. Current Status and Next Steps

The main technical obstacle under discussion is the final integration with the live Spotify API within the test_agent.py script. Preliminary tests have shown difficulty in creating API queries that effectively give song recommendations. This is mostly caused by 404 Not Found errors, which mean the agent's very particular musical profiles do not always have matching songs in the Spotify catalog. The code has been designed with a strong, multi-step fallback logic to manage these situations, starting from focused to more broad questions to guarantee that a song is almost always returned.

The following are the next immediate actions:

1. Run the final, updated test_agent. py script to debug and complete the API interaction logic, so guaranteeing that real songs are always suggested for the agent's generated profiles.
2. Perform a thorough qualitative analysis of the actual song recommendations, noting song sequences and assessing their fit for the intended emotional changes.

Bibliography

[1] Ardura Carnicero, A. (2023). *Development of a Music Recommender System to Promote Emotional Well-being*. Trabajo Fin de Grado, Universidad Politécnica de Madrid.

[2] Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

[3] Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

[4] Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207.

[5] Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences, 31*(5), 559–575.

[6] Eerola, T., & Vuoskoski, J. K. (2011). A review of music and emotion studies: Approaches, models, and current issues. *Psychology of Music, 39*(1), 3-37.

[7] Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion, 22*(4), 720-752.

[8] Gagnon, L., & Peretz, I. (2003). Mode and tempo relative contributions to "happy-sad" judgments in fast and slow music. *Cognition & Emotion, 17*(1), 25-40.

[9] Gabrielsson, A., & Lindström, E. (2010). The role of structure in the musical expression of emotions. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 367-400). Oxford University Press.

[10] Nummenmaa, L., Haroma, H., Hirvonen, J., Kangas, J., Kalliokoski, K., & Hietanen, J. K. (2023). Bodily maps of musical sensations across cultures. *Proceedings of the National Academy of Sciences, 120*(51), e2308859121.