Human Centered Evaluation of CLAP-Based Emotion Annotations for Therapeutic Music Using MTurk

Liam Stapley, Abhishek Karwankar, Daniel Stevens, & Matthew Louis Mauriello

University of Delaware

SENSIFY LAB

Motivation

- CLAP (Contrastive Language–Audio Pretraining) generates emotion tags by aligning audio & text embeddings.
- Enables large-scale music annotation without manual labeling.
- Outputs are weakly supervised and may not align with human emotional perception.
- **Our work**: Develops an MTurk-based framework to validate & refine CLAP-generated annotations.
  - Can help recommendations for music based on desired emotional profiles

Research Questions

① 🎵 **Can** contrastive learning models like CLAP reliably generate emotionally coherent tags for therapeutic music without explicit supervision?
② 👁️ **How** closely do CLAP-generated emotion annotations align with human emotional perception and interpretation?
③ 🔄 **Can** iterative refinement techniques, such as pseudo-labeling based on human feedback, enhance the validity and interpretability of weakly-labeled datasets?

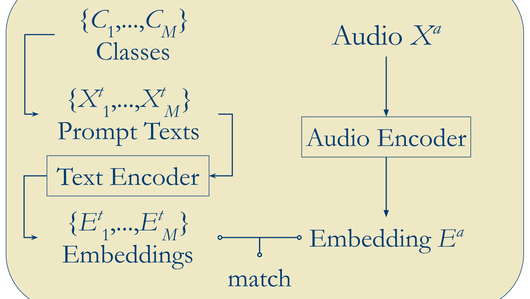
Methodology

Dataset Preparation

- **Source**: uCue's music library containing 10k+ possible song combinations.
- **Selection criteria**: Chose 432 tracks for diversity in emotion features.
- **Structure**: Modular arrangements with separate layers that can be added/removed.
- **Metadata**: Includes song IDs, modular layer info, and predefined emotion feature definit
- **Purpose**: Provide a representative subset of the full library for CLAP annotation and human validation.

Methodology cont.

CLAP Zero-Shot



GPT Description Generation

Prompt
- 3 most relevant features
- Feature definitions

Output
GPT generates a concise natural language description of the track's emotional character

Crowdsourced Evaluation (MTurk)

- Participants: Pre-qualified workers who passed a gold-standard clearance test.
- **Task Flow:**
  1. Listen to the full audio clip.
  2. Rank the top 3 emotion features.
     a. 8 features given with definition drop-down available
  3. Give 2-3 sentence description.
  4. Rate agreement with generated description.
     a. Only shown after "Next" button is clicked to prevent bias

MTurk Interface Workflow

Listen to Audio
- Full playback required (skip + mute prevention)

Rank Features
- Select and order top-3 emotion features

Write Description
- Write a 2-3 sentence natural language description

Rate AI Description
- Score GPT description on a Likert scale (-2 to 2)

Methodology cont.

Full Workflow

Dataset 🎵

CLAP Annotation
- Audio Encoder → Audio Embedding.
- Text Encoder → Feature Embeddings.
- Match via cosine similarity → Top-3 features.

GPT Description 🤖
Chat GPT generates natural language description from top-3 features + definitions.

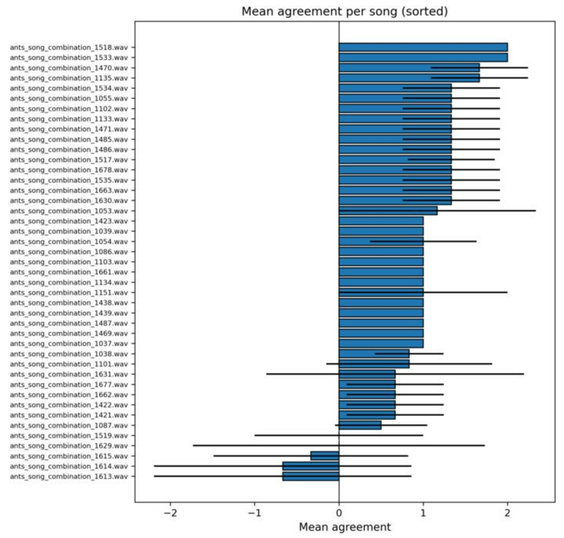
MTurk Evaluation 🧑‍💻
- Listen to full clip
- Rank top-3 features
- Provide human written description
- Rate GPT descriptions

SQL Database 🗄️
Stores timestamps, rankings, ratings, text feedback, and completion code

Analysis 📊
Agreement metrics, Tag overlap, Semantic similarity, Pseudo-label refinement

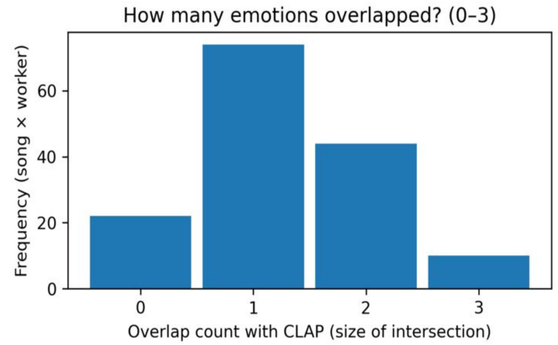
Preliminary Results

- **Agreement Score Analysis**: Mean & variance per song; identify high/low consensus cases.



Preliminary Results cont.

- **Top-3 Overlap**: Jaccard index & exact match counts between CLAP and human rankings.



Takeaway: CLAP typically shares exactly one label with a given worker.

Future Work

- Complete full dataset evaluation with MTurk.
- Apply pseudo-label refinement to improve annotation accuracy.
- Conduct demographic and contextual analyses of perception patterns.
  - **Description Similarity**: Embedding-based semantic similarity between GPT and human descriptions.
  - **Cluster Analysis**: Identify patterns of disagreement by emotion type or complexity.
  - **Pseudo-Label Refinement**: Use high-agreement samples to improve dataset accuracy.
- Publish validated therapeutic music dataset and open-source evaluation platform.
- Extend methodology to other domains, including environmental audio and affective speech.