Benchmark dataset for arousal-valence recognition

ISMIR 2022 BENGALURU

MusAV: A Dataset of Relative Arousal-Valence Annotations for Validation of Audio Models

Dmitry Bogdanov Xavier Lizarraga-Seijas Pablo Alonso-Jiménez Xavier Serra MTG, Universitat Pompeu Fabra

AV emotion recognition from audio

- **Our task**: Predict overall perceived emotion (arousal and valence, AV) of a music track from audio.
- **Problem**: Existing datasets are limited in coverage and do not represent a large variety of music available on commercial music platforms. There is no common benchmark dataset to compare models proposed by researchers and trained on different datasets.

Dataset	# tracks	Type	Source	Audio
EmoMusic	744 ft/exc	abs	MTurk	FMA
DEAM	1,802 ft/exc	abs	MTurk	FMA, Jamendo, MedleyDB
MuSe	41,021 exc	abs	Last.fm tags	Spotify (835 genres)
MER-TAFFC	900 exc	quad	manual	AllMusic
CCMED-WCMED	800 exc	rel	CrowdFlower	Classical music
EMusic	149 exc	rel	CrowdFlower	Experimental music

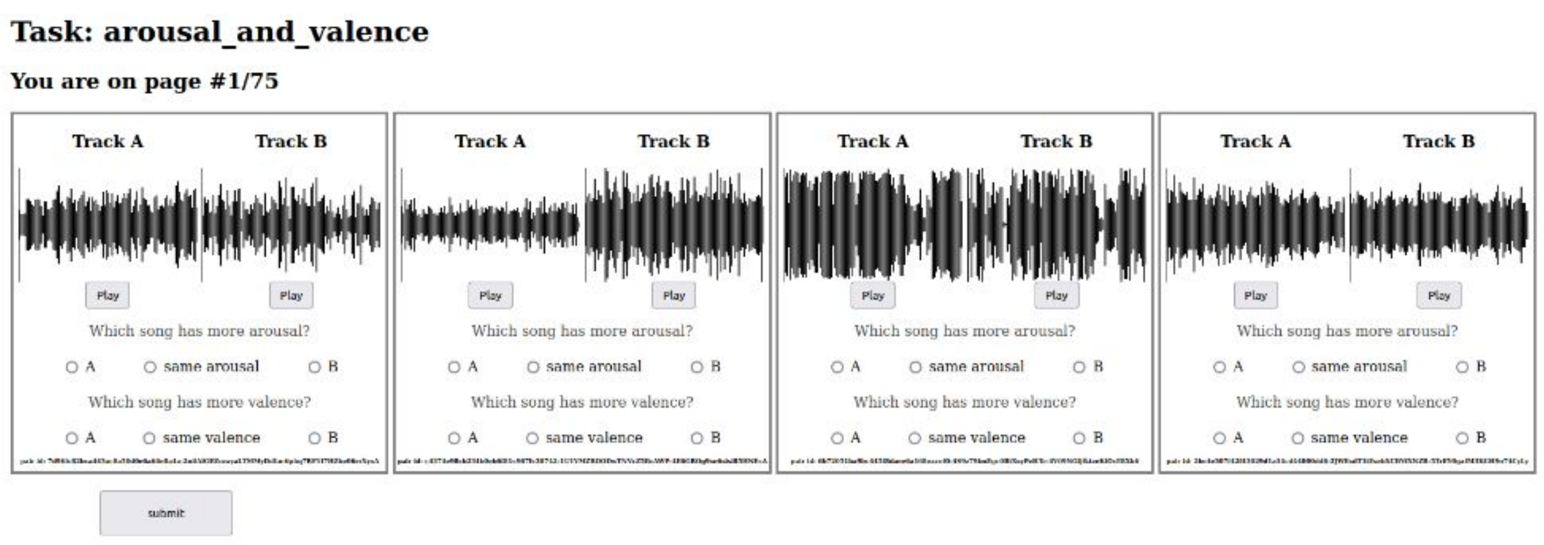
External validation with MusAV

- We trained and compared AV regression models built on 3 datasets with absolute AV annotations (EmoMusic, DEAM, MuSe) using 3 types of audio embeddings (EffNet-Discogs, MusiCNN-MSD, VGGish) [1-3].
- The downstream models are based on a fully connected layer with a linear activation function.
- In addition we used AV values provided by the Spotify API as an additional reference.
- All **pretrained models** are available as part of Essentia: https://essentia.upf.edu/models.html
- We evaluate our models on annotated pairs of tracks (e.g., song A has higher valence or arousal than song B), computing the percentage of pairs for which the model predictions correspond to the ground truth.

	Arousal				Valence			
	FA+MA	FA	FA+MA, CT	FA, CT	FA+MA	FA	FA+MA, CT	FA, CT
# track pairs	1413	950	716	502	1310	787	588	368
deam-effnet	72.28	75.44	72.60	74.84	61.59	63.38	63.91	65.51
deam-musicnn	78.81	81.04	76.92	78.41	59.75	61.98	62.33	62.90
deam-vggish	78.40	82.14	79.33	81.55	62.32	64.86	66.47	67.83
emomusic-effnet	82.57	86.55	84.75	87.61	71.29	75.41	73.77	78.55
emomusic-musicnn	85.61	89.21	84.78	87.63	**74.80**	**78.76**	76.53	80.29
emomusic-vggish	**86.42**	**90.30**	**86.86**	**89.73**	70.81	77.03	74.51	**81.16**
muse-effnet	59.92	60.99	59.00	62.11	62.14	63.78	61.54	64.35
muse-musicnn	63.96	66.59	64.84	68.55	67.72	70.77	69.03	71.01
muse-vggish	66.34	69.03	64.63	68.00	62.27	66.35	62.50	68.22
Spotify API	83.31	86.67	83.17	85.95	73.44	74.59	**77.51**	77.68

Dataset

- The dataset consists of **2,092 track previews** covering **1,404 genres**, with pairwise relative AV judgments by 20 annotators. We used Spotify API to preselect tracks and gather audio previews.
- Tracks are organized in triplets. For each pair in a triplet, 3 annotators voted on which song has higher arousal/valence using an **annotation tool** with loudness compensation.



- We gathered annotations for 7 annotation chunks with 100 triplets each, 20% **genre-triplets** (all tracks from the same genre), 80% **global-triplets** (tracks across different genres).
- We provide **ground truth** subsets of annotated track pairs based on different levels of agreement across annotators and triplet consistency (**full agreement** vs. **majority agreement** with/without **triplet consistency**).

Agreement	Arousal		Valence	
	# pairs	%	# pairs	%
FM+MA	1,448	69.4	1,341	64.3
FA	975	46.8	810	38.8
FM+MA, CT	738	35.4	606	29.1
FA, CT	519	24.9	381	18.3

- We observed fair to moderate agreement between annotators: ordinal Krippendorff's alpha of 0.48 for arousal and 0.39 for valence, consistent with previous studies.
- **License**: annotation metadata under CC BY-NC-SA 4.0. Audio previews available under request for non-commercial scientific research purposes only.

https://mtg.github.io/musav-dataset

[1] P. Alonso-Jiménez, X. Serra, and D. Bogdanov, "Music representation learning based on editorial metadata from Discogs," in International Society for Music Information Retrieval (ISMIR 2022), 2022.
[2] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," in International Society for Music Information Retrieval Conference (ISMIR 2019) Late Breaking Demo, 2019.
[3] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold et al., "CNN architectures for large-scale audio classification," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), 2017.

dmitry.bogdanov@upf.edu @di_bogdanov xavier.lizarraga@upf.edu @xavi_lizarraga pablo.alonso@upf.edu @pablo__alonso

This research was carried out under the project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.

upf. MTG Music Technology Group ESSENTIA MUSICAL AI_