3. ANNOTATION APPROACH

We follow a methodology that allows to avoid some of the limitations related to subjective annotation of arousal and valence with absolute values and instead consider comparative annotations on pairs of music tracks [25, 26]. Our main motivations are the following:

- Many practical applications are concerned not with the absolute AV values of songs, but with rankings songs according to these values. In such cases, relative annotations are convenient to validate ranking performance.
- For annotators, pairwise comparisons can be easier to understand and make decisions. They might require less effort than annotating with ranges of continuous values or Likert scales, which have more cognitive load [25,48].
- There is evidence that relative annotations have higher between-subject and within-subject agreement [25, 27].
- There are simple strategies to refine annotations according to the agreement between different annotators.
- Model evaluation can be a simple comparison of the ground-truth ordering of two tracks with the ordering according to the predicted AV values for all track pairs.
- Depending on how the models are trained, different models might predict AV values within different value ranges, which should be accounted for when comparing the performance metrics such as RMSE. Relative comparisons on pairs of tracks allows using simple common metrics that are compatible with all models.

To collect the dataset ground truth, we need an annotation tool able to reproduce pairs of music excerpts and collect the user's input. Such a tool should address potential sources of bias and simplify and speed up the annotation process. We define several requisites to accomplish this:

- We are interested in relative judgements about a pair of tracks (A and B) that can be formulated as the following question: "Which song has more *music property X*". The following choices are considered: *A*, *B*, or *same*. To minimize potential biases toward any of the choices, none of them should be selected by default.
- It may be difficult to maintain consistency when answering non-factual questions. Therefore, we consider that the interface should display multiple pairs in the same *page* to give the user an opportunity to improve coherence of their annotations before submitting the page.
- Loudness has a large impact in music perception. Higher loudness correlates to higher perceived arousal [29]. To minimize this effect, the annotation tool should include loudness normalization. Previous studies that proposed interfaces for AV annotations [16, 17, 25, 26] ignored this issue and did not include any normalization.

4. THE MUSAV DATASET

We created our dataset following the described methodology, using Spotify API as a source for music audio previews and genre metadata. Using Spotify allows us to access a wide range of music, while the 30-second previews it provides are sufficiently long to capture an overall perceived emotion with AV annotations.

4.1 Preparing the annotation pool

We collected a list of 5,716 genres from *everynoise.com* [5] which corresponds to the genre taxonomy of the Spotify API [6] and contains broad genres as well as specific subgenres. We then used the API's *Search* method to select random tracks for each genre. We generated multiple queries for each genre (using the genre tag, a wildcard search string starting with a random character, and a random market) and picked a random track from the list of the returned results for each query. This method allowed us to diversify music coverage and avoid popularity bias, downloading 17,574 track previews for 4,386 genres (up to 15 tracks per genre). All audio previews are 30-second long MP3 files with a 96 kbit/s bitrate. Each preview has a corresponding metadata file obtained with the API's *Get Track* method, including artist and album metadata and various audio analysis features.

We then analyzed the loudness of the audio previews to discard tracks with atypical levels, computing the integrated loudness in LUFS [49] of each track with the Essentia audio analysis library [50]. Based on the distribution of the obtained values, we kept tracks within the range between -20 and -5 LUFS, which represents the range of healthy loudness levels for the majority of mastered music. As a result, we reduced our pool to 15,979 tracks by 3,630 genres.

We organized the tracks into triplets with three pairwise comparisons each, allowing for additional inconsistency checks according to gathered relations within each triplet. We randomly assigned the tracks to two types of triplets: *genre-triplets* with all tracks sharing the same genre (one triplet per genre) and *global-triplets* containing tracks from various genres (the remaining tracks) to account for a use-case of distinguishing emotions within the same genre. All resulting 5,326 triplets contain unique tracks. We randomly split all generated triplets into annotation chunks, each one containing 100 triplets with 80% being global-triplets and 20% genre-triplets.

4.2 Annotation tool and process

We implemented a custom tool according to the requirements in Section 3. For each pairwise comparison (a pair), we used the wavsurfer-js [7] player to display a navigable representation of the waveforms. To prevent loudness bias, we normalize the songs to a common level of -20 LUFS. We computed the normalization factors from the LUFS values precomputed in the dataset preparation step, and converted them to linear gain units as expected by wavsurfer-js.

Our interface formulates two questions: "Which song has more arousal?" and "Which song has more va-

[5] `https://everynoise.com`
[6] `https://developer.spotify.com/documentation/web-api/reference`
[7] `https://wavesurfer-js.org/`

lence?". For both questions, the choices are *A*, *B*, or *same* arousal/valence. The tool shows a configurable number of pairs on each screen page (6 by default) and a submit button that stores the answers from the current page and renders the next one. We present the annotator with multiple pairs on the same page, as this can facilitate double-checking decisions made across pairs to minimize inconsistencies. Additionally, it simplifies navigation of the annotation interface, reducing the amount of necessary mouse movements and clicks. Finally, the tool has a page counter and displays each pair's ID to facilitate reporting any possible issues. The annotator outputs one JSON file per pair containing the answers to the two questions.

Figure 1 depicts the annotation tool. The source code for the annotation tool is publicly available online. [8] The tool is distributed as a Docker web application.

Due to the limitations on effort and availability of our annotators, we have proceeded with 7 annotation chunks which account to 2,100 tracks assigned to 700 triplets with 2,100 pairwise track comparisons. Overall, we gathered annotations from 20 participants, including authors' colleagues and students, with a background in music and technology. Each chunk was presented to three different annotators. Every annotator was given a single chunk, with an exception of one annotator who worked with two chunks. All annotators were instructed about the meaning of arousal and valence beforehand following their common definition [16, 42] and were asked to focus on perceived emotion [51]. Participants were aware of the subjectivity of the task and we encouraged to provide their subjective opinion. In total, we gathered 6,255 comparative arousal and valence judgments on pairs of tracks after discarding 15 pairs that the annotators reported having non-music tracks (speech) and duplicated tracks. These annotations involve 2,092 track previews by 1,404 genres.

4.3 Annotation agreement and consistency

By having multiple people annotate the same chunks of audio, we can measure the agreement between annotators. Computing ordinal Krippendorff's alpha, we obtained values of 0.48 for arousal and 0.39 for valence, which indicates a fair to moderate level of agreement, which is consistent with previous studies [26, 42, 46].

For building our ground truth for arousal and valence, we defined two types of agreement for pairwise comparisons of tracks by three different annotators. If all three annotators agreed on a pair of tracks with the same answer (*A-A-A*, *B-B-B*, or *same-same-same*) the annotations for this pair were considered to be in *full agreement*. If only two annotators agreed, we checked whether the third annotator was in a soft (e.g., *A-A-same*, *same-same-A*) or hard (*A-A-B* or *B-B-A*) disagreement. In the case of the former, we considered the annotations for the pair to be in *majority agreement*. Table 2 presents the agreement statistics for all gathered annotations.

In addition, we checked whether pairwise comparisons contradict each other within triplets (that is, whether they are geometrically inconsistent). For example, for three tracks X, Y, and Z forming a triplet, if $X > Y$ and $Y > Z$, but $X \leq Z$, such triplet and all its pairs are considered inconsistent. We considered triplets as consistent only if all constituent pairs had full or majority agreement in the annotations and no contradictions have been found.

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

Table 2. Number and percentage of annotated track pairs with different levels of annotator agreement and consistency for arousal and valence. FA+MA: pairs with full or majority agreement. FA: pairs with full agreement. CT: only pairs belonging to consistent triplets.

As a result, we generated different subsets of the annotations, with 69.4% and 64.3% of track pairs having at least some level of agreement and 24.9% and 18.3% passing the most strict conditions (pairs with full agreement, belonging to consistent triplets) for arousal and valence, accordingly.

Finally, as we had two types of triplets, we checked the effect of genre on the agreement rate: 67% and 61% of pairs in global-triplets had either full or majority agreement compared to 76% and 75% in the case of genre-triplets for arousal and valence, accordingly. This observation revealed that it was slightly easier to reach agreement on pairs of tracks coming from the same genre than from different genres.

4.4 Dataset contents

We provide the following contents as part of the dataset, available online: [9]

- Metadata for the entire annotation pool. [10] Each triplet is identified by a triplet ID and contains track Spotify IDs, triplet type (global-triplet or genre-triplet) and genre information.
- Split of the annotation pool into annotation chunks. [11]
- Raw comparative arousal and valence annotations on track pairs by anonymized annotators.
- Processed ground-truth annotations with different levels of agreement and consistency (full and major agreement with/without triplet consistency).
- Track audio previews and metadata gathered from the Spotify API for the annotated chunks. [12]
- Dataset metadata statistics (e.g., genre distribution).
- Scripts to reproduce the creation of the dataset.

[8] `https://github.com/MTG/musav-annotator`

[9] `https://mtg.github.io/musav-dataset`

[10] All annotation metadata is licensed under CC BY-NC-SA 4.0.

[11] It is possible to expand the dataset by annotating more chunks.

[12] Available under request for non-commercial scientific research purposes only. Any publication of results based on this data must cite Spotify API as the source of the data.

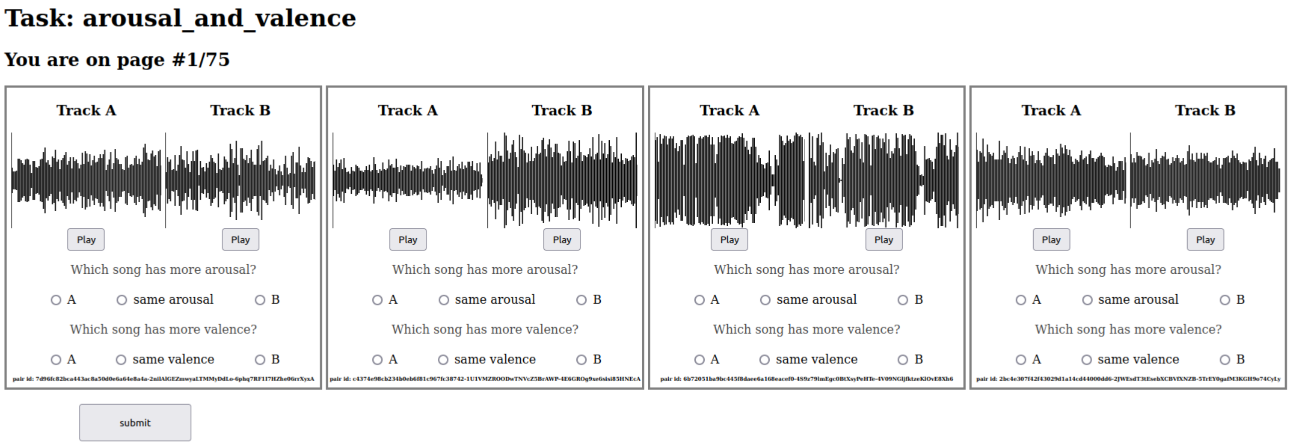

Figure 1. Screenshot of the annotation tool.

5. EXPERIMENTS

We demonstrate the dataset in use on the example of evaluating AV regression models based on audio embeddings.

5.1 Models

We created AV regression models based on three types of audio embeddings:

- **MusiCNN-MSD** (*musicnn*) is a music auto-tagging CNN with filter shapes motivated by music domain [52] trained on a subset of the Million Song Dataset. Its embedding layer has 200 units.
- **VGGish** (*vggish*) is a VGG architecture with an embedding layer of 128 units trained on audio from YouTube videos mapped to general purpose audio labels derived from their metadata [53]. The embeddings from this model were previously used for AV regression in combination with support vector regression [46].
- **EffNet-Discogs** (*effnet*) is an EfficientNet architecture trained to predict the music styles tags from Discogs. [13] The model produces 1280-dimensional embeddings and it is publicly available as part of Essentia models [54]. [14]

The embeddings are extracted on short one-, two-, and three-second audio excerpts for *vggish*, *effnet*, and *musicnn* models, accordingly. They are then used by the downstream regression models that we train. These models provide arousal-valence inference for each embedding vector, with a variable batch size with batch normalization. As Figure 2 depicts, we use a fully connected layer with a linear activation function, preceded by batch normalization and dropout. We also apply L1-L2 and L2 regularizers in the fully connected layer and dropout as regularization methods.

For training, we used three different datasets: DEAM, EmoMusic and MuSe, which we selected based on their music coverage as more appropriate for general use, with the goal to incorporate the resulting models as part of Essentia [24]. Each of them provides different audio excerpts (with 30-45 second duration) and arousal-valence values characterizing the overall emotion of each track. We extracted the embeddings with the pretrained models and used them as features. We followed a standard data splitting and loading strategy used in previous music classification publications [24]. To generate a train/test split stratified in terms of AV quadrants, we use Z-score normalization. However, we did not normalize the training data. The AV value range in all the datasets is from 1 to 9.

[13] https://blog.discogs.com/en/genres-and-styles

[14] https://essentia.upf.edu/models.html

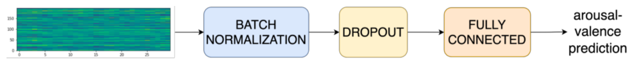

Figure 2. Arousal-valence backend model architecture.

Our models operate on short audio chunks and allow us to generate sequences of AV predictions over time with a new prediction every 1-3 seconds, depending on the receptive field of the embedding model used. Therefore, we compute the average of predictions on chunks to estimate the overall arousal/valence of a track.

In Table 3, we report Root Mean Square Error (RMSE) and Coefficient of Determination R^2 (R^2) commonly used for evaluation of regression models [17, 55, 56], obtained

	Arousal		Valence	
	R^2	RMSE	R^2	RMSE
deam-effnet	0.404	0.913	0.335	0.909
deam-musicnn	**0.417**	**0.894**	**0.400**	**0.818**
deam-vggish	0.396	0.963	0.344	0.963
emomusic-effnet	0.420	1.090	0.375	**0.948**
emomusic-musicnn	**0.451**	**1.030**	0.363	0.966
emomusic-vggish	0.429	1.037	**0.376**	0.973
muse-effnet	**0.143**	**1.148**	**0.089**	**1.581**
muse-musicnn	0.141	1.320	0.085	2.509
muse-vggish	**0.143**	**1.148**	0.085	1.584

Table 3. Evaluation metrics for the AV regression models on the held-out (testing) sets of the corresponding training datasets. The best values for each dataset are in bold.

	Arousal				Valence			
# track pairs	FA+MA 1413	FA 950	FA+MA, CT 716	FA, CT 502	FA+MA 1310	FA 787	FA+MA, CT 588	FA, CT 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

Table 4. External validation results on the proposed MusAV dataset (the percentage of track pairs with a correctly predicted ordering). FA+MA: pairs with full or majority agreement. FA: pairs with full agreement. CT: only pairs belonging to consistent triplets. The highest values are marked in bold. The top three AV models are marked in gray.

on the datasets used for training the models.

5.2 External validation on MusAV

We used our new dataset to validate the performance of the models. To this end, we assessed whether the ground truth ordering of track pairs coincided with the ordering according to the AV values predicted by the models. In addition, we also evaluated arousal and valence estimations provided by Spotify API and computed from audio as an additional reference. [15] This reference possibly represents a common state of the art in industrial systems. We ensured that our external validation set is independent of the datasets used for training the models: EmoMusic and DEAM contain non-commercial music unavailable on Spotify, while the intersection with MuSe, for which we also used Spotify track previews, includes 24 tracks that we filtered out for our evaluation.

For simplicity, we discarded all ground-truth annotations marking two songs as equivalent (13% and 15% of the ground-truth pairs with full or majority agreement in the case of arousal and valence, respectively). Thus, we focused only on examples with clear difference in arousal or valence. Table 4 presents the accuracy of the models in terms of the percentage of track pairs with correct ordering. We report the results on different subsets of the AV ground truth to demonstrate various evaluation possibilities.

5.3 Discussion

Having a new common ground truth for all models, our external validation shows the impact of the training dataset and embeddings.

Remarkably, models trained on the EmoMusic dataset perform the best for both arousal and valence regression. This is surprising, given that this dataset is smaller and less diverse than DEAM, which was derived from EmoMusic. On the other side, models based on MuSe have the worst performance in the case of arousal. Even though MuSe is the largest dataset in terms of size and coverage of commercially-available music, it appears to be too noisy to be able to train efficient models for arousal. Notably, it is the only dataset out of three relying on user-generated tags instead of explicit AV annotations, and the employed process for mapping tags to AV values might be inherently noisier.

Second, given a dataset, the choice of embedding model also matters. For example, the *effnet* embeddings trained on a large music style dataset appear to be inefficient for emotion recognition. The models based on them are consistently worst in the case of arousal (with all three datasets used for training) and valence (with EmoMusic and MuSe used for training). In turn, our validation reveals high performance of the *vggish* embeddings in many cases, possibly due to their generalization ability which was previously evidenced in literature [24]. This observation contradicts the results obtained in the respective held-out sets, where it did not have a remarkable performance overall.

Finally, in our validation, some of the considered AV models have performance competitive with an industrial reference. Still, all of the considered models only achieve up to 90% accuracy for arousal and 81% for valence. This is in line with evidence that predicting valence (as well as its annotation) is generally considered more complex than arousal [17, 42].

6. CONCLUSIONS

We present a new public dataset of relative AV annotations for validation of audio-based AV models. To build it, we employ a methodology that maximizes coverage in terms of genres to gather our annotation pool and allows to assess consistency of annotations on triplets of songs. The dataset is based on audio previews from Spotify API which allows validating performance on diverse types of commercially-available music. As an example, we train and evaluate AV regression models based on three common AV datasets and three types of pretrained audio embeddings and show how such a benchmarking can provide valuable complementary information about model performances. The resulting pretrained models are publicly available as part of Essentia models.

[15] We consider the "energy" descriptor in the Spotify API as arousal.

7. ACKNOWLEDGEMENTS

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. We also thank Juan Sebastián Gómez Cañón for his suggestions and all participating annotators.

8. REFERENCES

[1] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *International Society for Music Information Retrieval Conference (ISMIR 2010)*, 2010.

[2] Y.-H. Yang and H. H. Chen, *Music emotion recognition*. CRC Press, 2011.

[3] J. Grekow, *From content-based music emotion recognition to emotion maps of musical pieces*. Springer International Publishing, 2018.

[4] X. Hu, M. Bay, and J. S. Downie, "Creating a simplified music mood classification ground-truth set," in *International Conference on Music Information Retrieval (ISMIR 2007)*, 2007.

[5] C. Laurier, M. Sordo, J. Serra, and P. Herrera, "Music mood representations from social tags." in *International Society for Music Information Retrieval Conference (ISMIR 2009)*, 2009.

[6] X. Downie, C. Laurier, and M. Ehmann, "The 2007 MIREX audio mood classification task: Lessons learned," in *International Symposium on Music Information Retrieval (ISMIR 2008)*, 2008.

[7] C. Laurier and P. Herrera, "Audio music mood classification using support vector machine," in *International Conference on Music Information Retrieval (ISMIR 2007)*, 2007.

[8] K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music mood and theme classification-a hybrid approach." in *International Society for Music Information Retrieval Conference (ISMIR 2009)*, 2009.

[9] D. Bogdanov, A. Porter, P. Tovstogan, and M. Won, "MediaEval 2019: Emotion and theme recognition in music using Jamendo," in *MediaEval 2019 Workshop*, 2019.

[10] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[11] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[12] D. M. Greenberg, M. Kosinski, D. J. Stillwell, B. L. Monteiro, D. J. Levitin, and P. J. Rentfrow, "The song is you: Preferences for musical attribute dimensions reflect personality," *Social Psychological and Personality Science*, vol. 7, no. 6, pp. 597–605, 2016.

[13] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[14] J. A. Russell, "A circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[15] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models." in *International Society for Music Information Retrieval Conference (ISMIR 2009)*, 2009.

[16] M. Soleymani, A. Aljanaki, and Y. Yang, "DEAM: MediaEval database for emotional analysis in music," 2016. [Online]. Available: https://cvml.unige.ch/databases/DEAM/manual.pdf

[17] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *ACM International Workshop on Crowdsourcing for Multimedia (CrowdMM 2013)*, 2013.

[18] J. Grekow, "Music emotion maps in arousal-valence space," in *IFIP International Conference on Computer Information Systems and Industrial Management (CISIM 2016)*, 2016.

[19] J. Bai, J. Peng, J. Shi, D. Tang, Y. Wu, J. Li, and K. Luo, "Dimensional music emotion recognition by valence-arousal regression," in *IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC 2016)*, 2016.

[20] C. Akiki and M. Burghardt, "MuSe: The musical sentiment dataset," *Journal of Open Humanities Data*, vol. 7, no. 10, 2021.

[21] K. W. Cheuk, Y.-J. Luo, B. Balamurali, G. Roig, and D. Herremans, "Regression-based music emotion prediction using triplet neural networks," in *International joint conference on neural networks (IJCNN 2020)*, 2020.

[22] A. Livshin, "The importance of cross database evaluation in musical instrument sound classification," in *International Conference on Music Information Retrieval (ISMIR 2004)*, 2003.

[23] D. Bogdanov, A. Porter, P. Herrera, and X. Serra, "Cross-collection evaluation for music classification tasks," in *International Society for Music Information Retrieval Conference (ISMIR 2016)*, 2016.

[24] P. Alonso-Jiménez, D. Bogdanov, J. Pons, and X. Serra, "TensorFlow audio models in Essentia," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, 2020.

[25] Yi-Hsuan Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, 2011.

[26] J. Fan, K. Tatar, M. Thorogood, and P. Pasquier, "Ranking-based emotion recognition for experimental music," in *International Society for Music Information Retrieval Conference (ISMIR 2017)*, 2017.

[27] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013)*, 2013.

[28] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *IEEE International Conference on Affective Computing and Intelligent Interaction (ACII 2015)*, 2015.

[29] R. T. Dean, F. Bailes, and E. Schubert, "Acoustic intensity causes perceived changes in arousal levels in music: An experimental investigation," *PLOS One*, 2011.

[30] P. Kuppens, F. Tuerlinckx, J. A. Russell, and L. F. Barrett, "The relation between valence and arousal in subjective experience," *Psychological Bulletin*, vol. 139, no. 4, pp. 917–940, 2013.

[31] P. Kuppens, F. Tuerlinckx, M. Yik, P. Koval, J. Coosemans, K. J. Zeng, and J. A. Russell, "The relation between valence and arousal in subjective experience varies with personality and culture," *Journal of Personality*, vol. 85, no. 4, pp. 530–542, 2017.

[32] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021.

[33] J. Gómez-Cañón, E. Cano, Y. Yang, P. Herrera, and E. Gómez, "Let's agree to disagree: Consensus entropy active learning for personalized music emotion recognition," in *International Society for Music Information Retrieval Conference (ISMIR 2021)*, 2021.

[34] T. Li and M. Ogihara, "Detecting emotion in music," in *International Conference on Music Information Retrieval (ISMIR 2003)*, 2003.

[35] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5–18, 2005.

[36] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, 2014.

[37] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626.

[38] B. Bhattarai and J. Lee, "Automatic music mood detection using transfer learning and multilayer perceptron," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 19, no. 2, pp. 88–96, 2019.

[39] M. Soleymani, M. N. Caro, E. M. Schmidt, and Y.-H. Yang, "The MediaEval 2013 brave new task: Emotion in music," in *MediaEval 2013 Workshop*, 2013.

[40] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Emotion in Music Task at MediaEval 2014," in *MediaEval 2014 Workshop*, 2014.

[41] ——, "Emotion in Music Task at MediaEval 2015," in *MediaEval 2015 Workshop*, 2015.

[42] ——, "Developing a benchmark for emotional analysis of music," *PLOS One*, vol. 12, no. 3, p. e0173392, 2017.

[43] S. Zhao, Y. Li, X. Yao, W. Nie, P. Xu, J. Yang, and K. Keutzer, "Emotion-based end-to-end matching between image and music in valence-arousal space," in *ACM International Conference on Multimedia (MM 2020)*, 2020.

[44] H. Liu, Y. Fang, and Q. Huang, "Music emotion recognition using a variant of recurrent neural network," in *International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018)*, 2019.

[45] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[46] J. Fan, Y.-H. Yang, K. Dong, and P. Pasquier, "A comparative study of western and chinese classical music based on soundscape models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, 2020.

[47] J. Fan, M. Thorogood, and P. Pasquier, "Emo-soundscapes: A dataset for soundscape emotion recognition," in *IEEE International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, 2017.

[48] D. Yang and W.-S. Lee, "Disambiguating music emotion using software agents." in *International Conference on Music Information Retrieval (ISMIR 2004)*, 2004.

[49] European Broadcasting Union (EBU), "EBU Tech 3341. Loudness metering: 'EBU mode' metering to supplement loudness normalisation in accordance with EBU R 128."

[50] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "ESSENTIA: An audio analysis library for music information retrieval," in *International Society for Music Information Retrieval Conference (ISMIR 2013)*, 2013.

[51] N. F. Gutiérrez Páez, J. S. Gómez-Cañón, L. Porcaro, P. Santos, D. Hernández-Leo, and E. Gómez, "Emotion annotation of music: A citizen science approach," in *International Conference on Collaboration Technologies and Social Computing (CollabTech 2021)*, 2021.

[52] 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.

[53] 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.

[54] 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.

[55] C. J. Willmott, "On the validation of models," *Physical geography*, vol. 2, no. 2, pp. 184–194, 1981.

[56] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 2021.

WHAT IS MISSING IN DEEP MUSIC GENERATION? A STUDY OF REPETITION AND STRUCTURE IN POPULAR MUSIC

Shuqi Dai *
Carnegie Mellon University
shuqid@cs.cmu.edu

Huiran Yu *
Carnegie Mellon University
huiranyu@cs.cmu.edu

Roger B. Dannenberg
Carnegie Mellon University
rbd@cs.cmu.edu

ABSTRACT

Structure is one of the most essential aspects of music, and music structure is commonly indicated through repetition. However, the nature of repetition and structure in music is still not well understood, especially in the context of music generation, and much remains to be explored with Music Information Retrieval (MIR) techniques. Analyses of two popular music datasets (Chinese and American) illustrate important music construction principles: (1) structure exists at multiple hierarchical levels, (2) songs use repetition and limited vocabulary so that individual songs do not follow general statistics of song collections, (3) structure interacts with rhythm, melody, harmony and predictability, and (4) over the course of a song, repetition is not random, but follows a general trend as revealed by cross-entropy. These and other findings offer challenges as well as opportunities for deep-learning music generation and suggest new formal music criteria and evaluation methods. Music from recent music generation systems is analyzed and compared to human-composed music in our datasets, often revealing striking differences from a structural perspective.

1. INTRODUCTION

Structure is fundamental to music, as seen in the focus on form and analysis in music theory [1–3], music segmentation [4–6], structure analysis [7–9] and chorus detection [10] in MIR research, and recently in the attention given to long-term dependencies in music sequence generation using deep learning techniques [11, 12]. As a basic indicator of music structure, repetition contains important music information. Music relies heavily on repetition to create internal references, coherence and structure.

Unfortunately, the nature of repetition and structure in music is still not well understood, and much remains to be explored with music information retrieval techniques. For example, while music theory may suggest that songs have distinctive motives, our work quantifies and generalizes this notion. We will use "structure" to refer broadly to organizing principles in music, which are generally hierarchical and include sections, phrases and various kinds of patterns. A primary generator of structure is repetition, which includes not just music content within repeat signs but also approximate repetitions at different time scales.

In music generation, many researchers rely on deep learning models to capture music structure and organizing principles implicitly from data. However, repetition, especially long-term repetition structure, does not seem to emerge automatically in deep music generation. Deep learning is a promising direction, but such research should include evaluations where we can assess the successes and failures of new approaches. Moreover, some may argue that we do not need deep learning models to learn prevalent repetitions in music: we can produce repetition simply by generating phrases to the desired length and pasting them into a template. However, we will see that phrase structure, song structure, and other elements of music are intertwined, making this simple approach unable to reproduce characteristics of actual songs. Thus, we need a better understanding of repetition if we want machines to compose or even just listen to music in a more human way.

We aim to use formal models to explore music repetition and structure. By characterizing structural information in music, we can discover new principles of music organization and propose new challenges and evaluation strategies for music information retrieval and music generation.

An essential aspect of repetition structure is hierarchy. We use objective data analysis to support the existence and significance of multiple levels of hierarchy in popular music. We also present a number of results that show strong interactions between structure and pitch, rhythm, harmony, entropy and cross-entropy. Simply stated, structure can help predict pitch (or rhythm, harmony, etc.) and pitch (or rhythm, harmony, etc.) can help predict structure. These findings, in turn, challenge and inform research on deep learning to model hierarchical music structure.

Another important effect of repetition is that song-specific vocabulary of rhythm and pitch patterns is limited relative to what would be expected from the entire dataset. This vocabulary serves both to unify multiple phrases within a song and distinguish songs from others.

The main contribution of this work is a better understanding of the nature of repetition in popular music. We will see that repetition exists in many forms and at different levels of hierarchy. We offer ways to quantify music repetition structure, especially as it relates to pitch and

* The first two authors have equal contribution.

Attribution: S. Dai, H. Yu, and R. B. Dannenberg, "What is missing in deep music generation? A study of repetition and structure in popular music", in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.,* Bengaluru, India, 2022.

rhythm, often by measuring entropy or cross-entropy. We also reveal striking differences from a structural perspective through case studies on recent deep music generation models. These and other findings offer challenges as well as opportunities for deep-learning music generation and suggest new formal music criteria and evaluation methods.

After describing related work in the next section, we discuss music repetition and structure in Section 3, how it relates to deep music generation in Section 4, and we present conclusions in Section 5.

2. RELATED WORK

Repetition is a key element of music structure. Repetition is one of the three commonly used principles for segmenting music, along with novelty at segment boundaries and homogeneity within segments [7]. People have developed a variety of segmentation and section detection methods based on repetition with acoustic features [4, 10]. Repetition becomes especially useful in segmenting symbolic music or lead sheet representations where timbre and texture may be lacking [13].

Repetition also plays an important role in music expectation and prediction [14, 15]. Studies of repetition and structure are common in Music Psychology [16]. For example, listening experiments with reordered Classical and Popular music have shown that listeners are rather insensitive to restructuring, but these results are subtle and somewhat ambiguous [17]. Music form and structure, including repetition, is also a major focus of Music Theory [1,18,19].

There are many deep learning models for music generation [11, 20–23], however, capturing repetition and long-term dependencies in music still remains a challenge. One mainstream approach is to model distribution of music via an intermediate representation (embedding), such as Variational Auto-Encoders (VAE) [20,24], Generative Adversarial Networks (GANs) [25] and Contrastive Learning [24, 26]. Due to their fixed-length representation and short-length output, it is difficult to exhibit long-term structure. Another popular trend is to use sequential models such as LSTMs and Transformers [11, 22, 27] to generate longer music sequences, but they still struggle to generate repetition and coherent structure on long-term time scales. Some recent work introduces explicit structure planning for music generation, which shows that using structure information leads to better musicality [12, 28, 29].

Current evaluation methods for music generation rarely consider repetition and structure. Deep music generation systems [11,30] use objective metrics such as negative log-likelihood, cross-entropy and prediction accuracy to compare generated music with ground-truth human-composed music. But these metrics do not precisely correspond to human perception and are not reliable for musicality. Another trend is using domain-knowledge [31] and musical features [32–35] such as pitch class, pitch intervals, and rhythm density to evaluate music statistically. However, most of them ignore even short patterns, and none evaluate music structure. In contrast, we offer quantitative and objective methods to evaluate music repetition and structure.

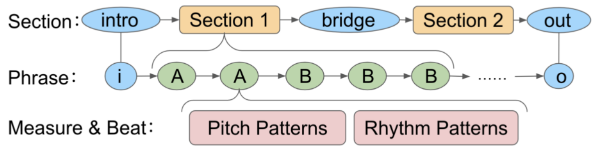

Figure 1: Structure hierarchy in pop music.

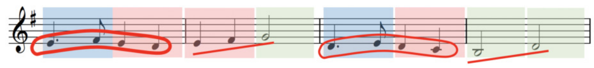

Figure 2: Repeated motives in a phrase in *Yankee Doodle*.

3. REPETITION AND STRUCTURE IN MUSIC

We are interested in three main problems concerning repetition and structure in music: (a) How are repetition and structure organized hierarchically? (b) How do different levels of structure interplay with other music elements? (c) How does repetition play out over time?

Unlike traditional music theory with case-by-case human analysis, we explore these problems in a data-driven approach. For training and testing, we use a Chinese pop song dataset POP909 [36], which has 909 pop song performances in MIDI, and an American pop song dataset PDSA [37], in MusicXML, which has 348 American pop songs originating from 1580 to 1924. We use only songs in 4/4 time to simplify analysis.

3.1 Repetition and Structure Hierarchy

Music structure is hierarchical. It contains multiple levels of repetitions, ranging from low-level pitch and rhythm motives (patterns) to higher-level phrases (analogous to sentences) and sections (analogous to paragraphs). We describe these in more detail and use statistical and machine learning findings in the data to explore their significance.

3.1.1 Phrases and Sections

Researchers [13] found two levels of structure in POP909: *sections* and *phrases*. Phrases were identified by searching (automatically) for approximate repetitions of sequences of 4 or more measures. Non-unique phrases (those that match other phrases) often occur in sequences such as "AABBB" in Figure 1 called *sections*, which are by definition separated by non-repeated or non-melodic phrases. For example, Figure 1 has an intro, two sections connected by a bridge transition, and an outro, which is a typical pop song structure. Here "A" can be a verse, and "B" can be a chorus phrase. These phrase and section levels of structure were found to be predictive of aspects of pitch, rhythm and harmony, which shows that these two levels have significance for composition, probably for perceptual reasons.

3.1.2 Repetition Below the Phrase Level

There is at least another level of repetition below phrases (Figure 1). For example, in the first 8-measure phrase of the chorus in *Yankee Doodle* (Figure 2), the first and second half repeat elements of rhythm and interval. The colored boxes show repeated rhythm patterns, and the red lines point out repeated pitch contours. We want to assess whether this kind of repetition is common.