

Expressivity-aware Music Performance Retrieval using Mid-level Perceptual Features and Emotion Word Embeddings

Shreyan Chowdhury

Institute of Computational Perception
Johannes Kepler University Linz, Austria
shreyan0311@gmail.com

ABSTRACT

This paper explores a specific sub-task of cross-modal music retrieval. We consider the delicate task of retrieving a performance or rendition of a musical piece based on a description of its style, expressive character, or emotion from a set of different performances of the same piece. We observe that a general purpose cross-modal system trained to learn a common text-audio embedding space does not yield optimal results for this task. By introducing two changes – one each to the text encoder and the audio encoder – we demonstrate improved performance on a dataset of piano performances and associated free-text descriptions. On the text side, we use emotion-enriched word embeddings (EWE) and on the audio side, we extract mid-level perceptual features instead of generic audio embeddings. Our results highlight the effectiveness of mid-level perceptual features learnt from music and emotion enriched word embeddings learnt from emotion-labelled text in capturing musical expression in a cross-modal setting. Additionally, our interpretable mid-level features provide a route for introducing explainability in the retrieval and downstream recommendation processes.

ACM Reference Format:

Shreyan Chowdhury and Gerhard Widmer. 2023. Expressivity-aware Music Performance Retrieval using Mid-level Perceptual Features and Emotion Word Embeddings. In *Forum for Information Retrieval Evaluation (FIRE 2023), December 15–18, 2023, Panjim, India*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3632754.3632761>

1 INTRODUCTION

Music performance involves an interplay between a composer, a performer, and a listener. The performer imbues a composition with their own expression and style, and the listener perceives the emotion or mood being communicated through the music by the composer and the performer [19]. Thus the *expressive quality* of performed music becomes an important attribute through which a musical performance is characterised [4, 15]. The present paper provides a step towards building music retrieval and recommendation systems that are sensitive to this expressive quality of music.

We aim to develop a system for retrieving a desired performance of a musical piece from a set of different performances based on a description of its style, expressive character, or emotion. This is an important problem when dealing with certain genres of music

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

FIRE 2023, December 15–18, 2023, Panjim, India

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1632-4/23/12.

<https://doi.org/10.1145/3632754.3632761>

Gerhard Widmer

Institute of Computational Perception
Johannes Kepler University Linz, Austria
gerhard.widmer@jku.at

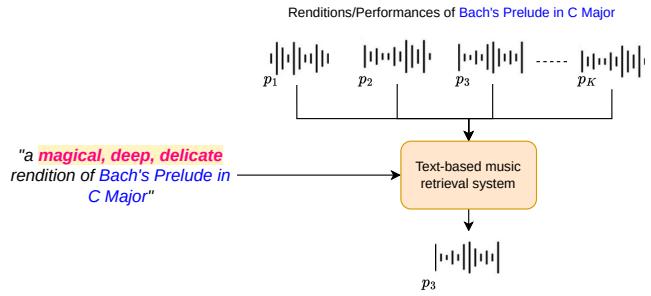


Figure 1: A system for retrieving the best-matching performance of a musical piece based on a text description of its expressive character.

such as Western classical music, where we typically have several renditions of compositions performed by different artists¹, and possibly very distinguishing listeners. These interpretations, while all representing the same musical content (piece), could vary subtly or vastly in their expressive character, and traditional recommendation and retrieval algorithms are not designed to be sensitive to such performance characteristics.

In our work, we look at text-based music retrieval. Formally, given a set of K performances $P = \{p_1, p_2, \dots, p_K\}$ of a musical piece and a query $Q = \{w_1, w_2, \dots, w_N\}$ comprised of N “descriptive” words, we would like to retrieve the performance (say p_3 , as shown in Figure 1) that we believe a general audience might consider matches best with the description. This is a highly subjective task. We thus evaluate our models on the *Con Espressione* dataset [7], which is a dataset comprised of different piano performances, by famous pianists, of a number of classical piano pieces, and associated free-text descriptions provided by contributors through an online questionnaire (more details in Section 3). In this dataset, we have on an average of five performances per piece, which thereby constitute the search space for each query. The assumption here is that we already know the piece (and thus the search space for that piece) for each query, and we want to retrieve the closest matching performance in terms of its expressive character. Inferring the piece itself from the query, while a necessary component of an end-to-end retrieval system, is not the focus of the present paper. The reader is encouraged to look at works on music search and retrieval such as [26] for more details on this topic.

The subjective nature of the task also motivated us to consider human-oriented, perceptually relevant descriptors for representing expressive qualities extracted from audio recordings. In particular,

¹As stated by a major music streaming platform, “Classical music is different. It has longer and more detailed titles, multiple artists for each work, and hundreds of recordings of well-known pieces” [1].

we focus on so-called *Mid-level Perceptual Features* – relatively high-level musical qualities that are considered to be perceptually important [5]; in our work, these are learnt from human annotations. Previous research has shown these features to have the capacity to predict musical emotions as well as to disentangle different performances based on emotion [7–9].

Furthermore, on the text side, we note the inefficiency of traditional word embeddings (Word2Vec, GloVe) to capture the intended emotion in descriptions of musical performances. Such language representation models carry the inductive bias that words used in the same context tend to possess similar meanings, thus resulting in emotionally dissimilar words like *happy* and *sad* having close proximity in the representation space, due to them often occurring in similar contexts in the training data. To derive correct emotional meanings from our text queries, we experiment with emotion-enriched word embeddings from Agrawal et al. [3] and find that they improve retrieval results significantly, particularly when combined with mid-level features on the audio side.

2 RELATED WORK

Our work sits in the broader area of cross-modal retrieval. This is an area of active research in general, and is of great significance to music information retrieval [14, 20, 22]. Language-based audio retrieval is currently witnessing much interest from the research community as can be seen from active participation in competitions like DCASE [12]. In this section, we look at some recent work in text-based retrieval for music.

Audio-Language Learning for Music: Text-based audio retrieval is typically done by learning a common embedding space of aligned audio and text embeddings. In “MusCALL” [21], this is done using a multimodal contrastive learning (MCL) approach. Two encoders, $f_a(\cdot)$ for audio and $f_t(\cdot)$ for text, are learnt such that for any audio-text pair (a_i, t_i) , the resulting embeddings $z_{a,i} = f_a(a_i)$ and $z_{t,i} = f_t(t_i)$ lie close in the joint embedding space [24]. A ResNet-50 [17] is used for the audio backbone, and a downsized-Transformer [27] for the text backbone. They use 250k audio-text pairs from a production music library as the dataset. In “MuLan” [18] a much larger dataset of 44m recordings and weakly-associated, free-form text annotations is used. They use a pre-trained BERT [13] as their text encoder and experiment with two different audio encoders – ResNet-50 and Audio Spectrogram Transformer [16].

A more recent work is “Music and Text Representation” learning (MTR) [14]. It lays out an investigation into effective design choices for universal representation learning for text-to-music retrieval systems. This work uses a set of 500k music-text pairs. They use a modified version of the Music Tagging Transformer [28] as the audio encoder, and two different text encoders – pre-trained GloVe [23] and pre-trained BERT. We were able to obtain the trained models for this work, which we use as a baseline for our work.

Emotion Embedding Spaces: In [30], the authors propose “emotion embeddings” for retrieval of musical pieces that match the emotional characteristic of stories. They use a ResNet model [29] as the audio backbone and a pre-trained DistilBERT [25] as the text backbone. However, they do not use natural language in the audio-text pairs in their experiments; rather the text component comes from the labels or tags associated with the audio clips, which

are mapped to the embedding space using a Word2Vec model or emotion lexicons.

2.1 Music-Text Representation (MTR)

We consider the Music-Text Representation (MTR) model by Doh et al. [14] as the baseline system. It is a cross-modal retrieval system that projects audio and text representations onto a common embedding space and reduces the distance between paired vectors during training. It is trained on 500k audio-text pairs with the text formed by concatenating tags from different sources. This data is a subset of the Million Song Dataset [6] and contains songs from a mix of different genres. We use weights of the best version of their system according to the results on their paper, which is a contrastive model type with a BERT text encoder and stochastically sampled text representations.

3 DATASETS

The Con Espressione Dataset [7] consists of recordings of 9 piano pieces played by different famous pianists (making a total of 45 performances), and associated free-text descriptions for each performance, collected from a large number of listeners. The study participants were asked to describe the *expressive character* of each performance. Typical characterisations in the dataset are adjectives like “cold”, “playful”, “dynamic”, “passionate”, but also more complex phrases such as “controlled with speed”, “smooth tempo variation”, “emotional with dynamics” etc. In this work, we use the aggregated answers for each performance. That is, all words or phrases used by different participants for a performance are concatenated into a single text description of the performance.

The Mid-level Perceptual Features Dataset [5] consists of 5k song snippets of 15 seconds each annotated according to seven mid-level descriptors: *melodiousness*, *articulation*, *rhythm stability*, *rhythm complexity*, *dissonance*, *tonal stability*, and *modality* (or ‘*minorness*’). The ratings for the seven mid-level perceptual features were collected through crowd-sourcing. We use a model trained on this dataset as our audio encoder, and we refer to this trained model as the “mid-level model”, or “mid-level encoder” in the following paragraphs. We add *onset density* as an additional feature according to [11] (this gives the best results in our case).

Additional Datasets: MusicCaps and Pitchfork Track Reviews. In order to effectively train a model, we require a substantial amount of audio-text paired data, and unfortunately, the Con Espressione dataset is insufficient in size for this particular task. To expand our dataset, we look towards two sources: the MusicCaps dataset, and a trusted and well-known music review website, Pitchfork (www.pitchfork.com).

The MusicCaps dataset [2] consists of 5.5k music clips from diverse genres with paired text descriptions

4 APPROACH

We define performance retrieval as the task of retrieving a particular musical performance from a set of K performances $P = \{p_1, p_2, \dots, p_K\}$ of a musical piece such that the returned performance best matches the query $Q = \{w_1, w_2, \dots, w_N\}$ comprised of N words. In our case, the query is a description of the desired expressive character of a performance provided in the form of text.

For each query, the piece is known, so the search space for the system is the set of performances of that piece. This is arguably a hard task; the retrieval system will have to be sensitive to very subtle musical differences. In our experiments, we will consider a retrieval result "correct" if the output of the system ranks the performance corresponding to the input text the highest.

We take the basic framework of Music-Text Representation (MTR) by Doh et al. [14] and modify the text and audio encoders. We hypothesise that two kinds of modifications might improve the model's effectiveness for expressivity-aware performance retrieval. First we need an audio encoder trained to extract features that are more attentive to the expressive qualities in music. Mid-level perceptual features [5] have a significant capacity to capture such musical qualities and hold good predictive power for music emotion [8]. We thus replace the audio encoder in MTR with a mid-level feature model. This pre-trained model takes audio spectrograms as inputs and outputs seven mid-level features (seven real-valued scalars). To this we add an eighth feature: *onset density* [11], intended to model the 'perceptual speed' of a performance. The assumption here is that points close together in the space spanned by the mid-level features are similar in their expressive quality. Second, we investigate if the system works better with a text encoder trained using emotion labels. While the text encoder in MTR is a state-of-the-art BERT sentence model, in our task, the sentence structure has less of a consequence than obtaining accurate representations of the descriptive words. We thus experiment with an emotion-enriched word embedding model (EWE) [3].

4.1 Mid-level Features Audio Encoder and Emotion Enriched Text Encoder

As reasoned earlier in this paper, we propose replacing the audio encoder in MTR with a trained mid-level feature model (f_m in Figure 2). The input to this model is an audio spectrogram, the output is a vector of 8 mid-level features (including onset density). This model is domain-adapted for piano music, since that is our domain of interest for this paper. As shown in [10], domain adaptation can improve mid-level prediction for piano audio without compromising the performance for other styles of music.

The text encoder is also replaced by an emotion enriched word embedding (EWE) model (g_{EWE} in Figure 2), which outputs an embedding of dimension 300 for each word. For a set of descriptive words, we take the resultant vector by adding all individual embeddings element-wise.

Now since these two encoders are pre-trained, they do not project to a common embedding space. Hence we need a model to project the outputs of the encoders to a common space for enabling cross-modal retrieval. We choose to project the text embeddings onto the mid-level feature space using a linear model h . Preserving the mid-level feature predictions has the additional advantage of providing explainable insight into the retrieval process, due to the dimensions possessing interpretable musical meanings. This is described in the next section. For h , a simple linear regression model proved sufficient. In some cases (see Table 2), transforming the text embeddings with principal component analysis (PCA) improved the results.

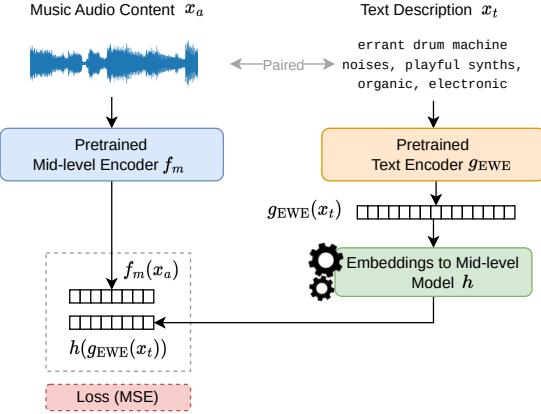


Figure 2: In our system, the audio and text encoders of a MTR model [14] are replaced by a mid-level feature model and emotion enriched word representation model respectively.

4.2 Experiments and Results

We perform piece-wise cross-validation over the Con Espressione dataset to fully utilise the available data, i.e., in each run the test set is the set of all performances of a piece in the dataset. The rest of the dataset is the train set for that run. This results in 9-fold cross-validation as we have 9 pieces in the dataset. For retrieval, we use the cosine similarity in the mid-level feature space:

$$\text{cosine_similarity}(\mathbf{m}_1, \mathbf{m}_2) = \frac{\mathbf{m}_1 \cdot \mathbf{m}_2}{\|\mathbf{m}_1\| \cdot \|\mathbf{m}_2\|} \quad (1)$$

We use top- k ratio and Mean Reciprocal Rank (MRR) as evaluation metrics. Top- k ratio (ranges from 0 to 1, with higher scores indicating better retrieval performance) is defined as the number of queries for which the correct performance has a rank equal to or better than k among all performances of the piece for which the query is made. We use $k = 1$ and $k = 2$.

Mean Reciprocal Rank (MRR) is defined as the average of the reciprocal rank of the correct item retrieved over a set of queries:

$$\text{MRR} = \frac{1}{|M|} \sum_{i=1}^{|M|} \frac{1}{\text{rank}_i} \quad (2)$$

where $|M|$ is the total number of queries, and rank_i is the rank of the correct item in the result list for the i -th query. MRR ranges from 0 to 1, with higher scores indicating better performance of the ranking algorithm.

We also investigate the impact of augmenting the train set with additional data from the MusicCaps and Pitchfork datasets (see Table 2).

Despite reasonable performance of traditional models on the general purpose task of music retrieval using text input, there remains much room for improvement for expressivity-aware music performance retrieval. We observe (see Table 1, second row) that for the MTR model the mean reciprocal rank for performance retrieval on the Con Espressione dataset is only 0.46, which means the average rank of the correct performance is greater than 2 (on sets of, on average, 5 performances per piece); in effect, this is similar to what

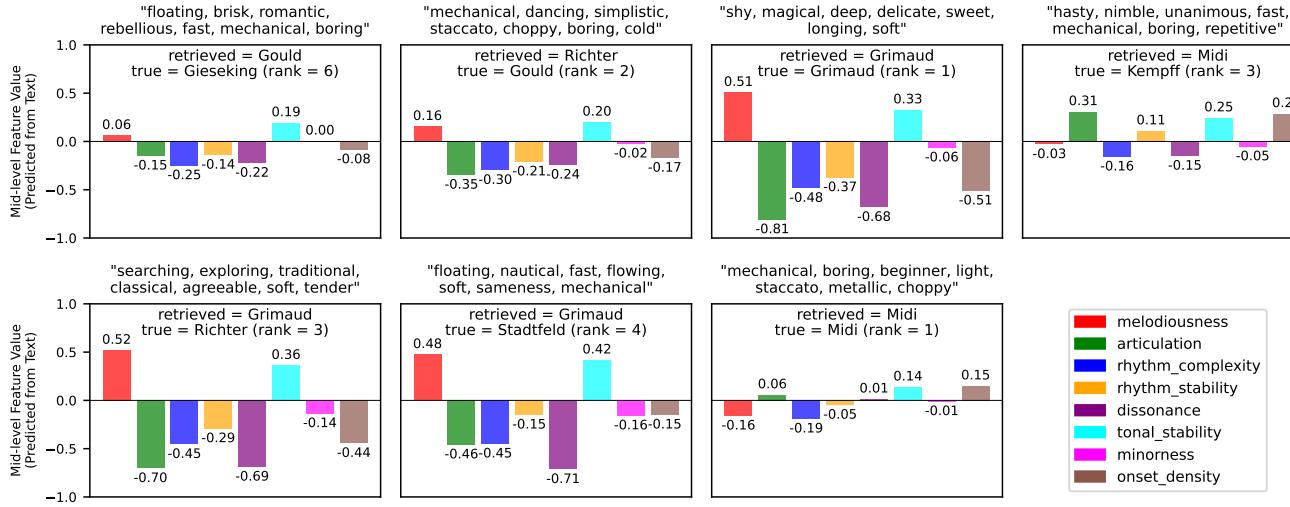


Figure 3: Seven performances of Bach’s C Major Prelude, with human textual characterisation and corresponding mid-level feature values predicted by mapping $h(gEWE)$. Text and feature values relate to the performance identified by “true = ...”

Table 1: Retrieval results using different audio and text encoders.
The values here are for a model trained with both Pitchfork and MusicCaps augmentation and no PCA on text embeddings (see Table 2). For the random baseline, performances are chosen at random.

Text Enc	Audio Enc	Top-1	Top-2	MRR
(Random Baseline)		0.18	0.37	0.44
MTR	MTR	0.20	0.42	0.46
MTR	Mid-level	0.22	0.53	0.50
EWE	MTR	0.22	0.42	0.47
EWE	Mid-level	0.38	0.64	0.61

we obtain with random guessing (Table 1, 1st row). Moreover, only 20% of the queries return the correct performance as the top result.

On the other hand, our model with emotion word embeddings (EWE) and mid-level features returns the correct performance for

Table 2: Effect of data augmentation and PCA on retrieval based on EWE embeddings and Mid-level features

Augmentation	Text Emb	Metrics		
		Top-1	Top-2	MRR
X	X	X	0.22	0.46
X	X	✓	0.29	0.49
✓	X	X	0.24	0.44
✓	X	✓	0.40	0.58
X	✓	X	0.33	0.57
X	✓	✓	0.27	0.53
✓	✓	X	0.38	0.64
✓	✓	✓	0.36	0.56

38% of queries, with an MRR value of 0.61 (Table 1, last row), meaning the average rank of the correct performance is about 1.63. While our method shows significant improvement over the baseline, it is important to note that this is a highly subjective task and the dataset we have at hand is not large. The main objective of this paper is to provide a proof-of-concept that models trained using domain-specific perceptual features can lend a significant advantage in cross-modal retrieval applications.

From Table 2, we see that augmenting with additional data has a minor but positive effect on the results whereas PCA tends to improve performance for non-augmented datasets. For the augmented case, PCA actually worsened the performance.

5 MID-LEVEL FEATURES AS EXPLANATIONS

It is instructive to look at an individual example. The mid-level features that our model predicts from the audio recordings, and which it uses to establish a relationship to textual descriptions, can be viewed as a kind of *explanation* of the model’s retrieval choice, pointing to musical qualities in the performance that may have influenced the decision. In the example shown in Figure 3, the piece in question is J.S.Bach’s Praeludium in C major from his *Well-Tempered Clavier* (Book I), for which our database contains seven different recordings, by Walter Giesecking, Glenn Gould, Hélène Grimaud, Wilhelm Kempff, Sviatoslav Richter, Martin Stadtfeld, and one completely expressionless, mechanical MIDI performance derived from the score. Given the query “shy, magical, deep, delicate, …”, the system returns the Grimaud performance (which is the “correct” one, as these descriptions were indeed associated with her performance by the human annotators); “mechanical, boring, beginner” returns, again correctly, the MIDI rendering. (The baseline MTR model also gets the Grimaud right, but returns Sviatoslav Richter in response to “mechanical, boring, beginner”). The mid-level feature profiles predicted from these text queries by our model (i.e., its ‘translation’

of free text into mid-level feature space) explain its decisions quite well and conform both to our intuition, and to what we hear in the recordings. In particular, Grimaud is described as having high "melodiousness" and very low "articulation" whereas the MIDI rendering is characterised by rather neutral values throughout, and negative rather than positive "melodiousness". Such high-level musical descriptions could be useful as explanations in a real music search and recommendation application.

6 CONCLUSION

We present an experimental demonstration showing that mid-level perceptual features and emotion enriched text embeddings are useful in capturing some of the expressive musical character in audio recordings and relating them to descriptive text. We see significant improvements in all three metrics (particularly, the number of times the correct performance was retrieved almost doubled) and we note the effect of modifying *both* the audio and the text encoders. While this is not a large-scale study, our results point towards the importance of perceptually-driven features and emotion-aware models in music retrieval and consequently in music recommendation. Features that are interpretable and musically meaningful are also crucial for explainability and provide a route for training downstream interpretable models of music retrieval and recommendation.

ACKNOWLEDGMENTS

This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreements No 670035 ("Con Espressione") and 101019375 ("Whither Music?").

REFERENCES

- [1] 2023. Apple Music Classical. <https://support.apple.com/en-us/HT213415>
- [2] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. 2023. MusicLM: Generating Music From Text. *ArXiv* abs/2301.11325 (2023).
- [3] Ameeta Agrawal, Aijun An, and Manos Papagelis. 2018. Learning emotion-enriched word representations. In *Proceedings of the 27th international conference on computational linguistics*. 950–961.
- [4] Jessica Akkermans, Renee Schapiro, Daniel Müllensiefen, Kelly Jakubowski, Daniel Shanahan, David Baker, Veronika Busch, Kai Lothwesen, Paul Elvers, Timo Fischinger, et al. 2019. Decoding emotions in expressive music performances: A multi-lab replication and extension study. *Cognition and Emotion* 33, 6 (2019), 1099–1118.
- [5] Anna Aljanaki and Mohammad Soleymani. 2018. A Data-driven Approach to Mid-level Perceptual Musical Feature Modeling. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*. Paris, France, 615–621.
- [6] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011*. Miami, Florida.
- [7] Carlos Cancino-Chacón, Silvan David Peter, Shreyan Chowdhury, Anna Aljanaki, and Gerhard Widmer. 2020. On the Characterization of Expressive Performance in Classical Music: First Results of the Con Espressione Game. In *Proceedings of the 21st International Society for Music Information Retrieval Conference, ISMIR 2020*. Online.
- [8] Shreyan Chowdhury. 2022. *Modelling Emotional Expression in Music Using Interpretable and Transferable Perceptual Features*. Ph.D. Dissertation. Johannes Kepler University Linz, Austria.
- [9] Shreyan Chowdhury and Gerhard Widmer. 2021. On Perceived Emotion in Expressive Piano Performance: Further Experimental Evidence for the Relevance of Mid-level Perceptual Features. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021*. Online.
- [10] Shreyan Chowdhury and Gerhard Widmer. 2021. Towards Explaining Expressive Qualities in Piano Recordings: Transfer of Explanatory Features Via Acoustic Domain Adaptation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*. Online, 561–565. <https://doi.org/10.1109/ICASSP39728.2021.9413638>
- [11] Shreyan Chowdhury and Gerhard Widmer. 2023. Decoding and Visualising Intended Emotion in an Expressive Piano Performance. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Late-breaking Demo Session*. Bangalore, India.
- [12] DCASE Challenge. 2023. Language-based audio retrieval. <https://dcase.community/challenge2023/task-language-based-audio-retrieval>
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. New Orleans, Louisiana, USA.
- [14] SeungHeon Doh, Minz Won, Keunwoo Choi, and Juhan Nam. 2023. Toward Universal Text-To-Music Retrieval. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. <https://doi.org/10.1109/ICASSP49357.2023.10094670>
- [15] Alf Gabrielsson and Patrik N. Juslin. 1996. Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience. *Psychology of Music* 24, 1 (1996), 68–91. <https://doi.org/10.1177/0305735696241007>
- [16] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proceedings of Interspeech 2021*. Brno, Czech Republic, 571–575.
- [17] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 770–778.
- [18] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. MuLan: A Joint Embedding of Music Audio and Natural Language. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022*. Bangalore, India.
- [19] Patrik N Juslin. 2013. What does music express? Basic emotions and beyond. *Frontiers in psychology* 4 (2013), 596.
- [20] Bochen Li and Aparna Kumar. 2019. Query by Video: Cross-modal Music Retrieval. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*. Delft, The Netherlands, 604–611.
- [21] Ilaria Mancò, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive Audio-Language Learning for Music. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022*. Bangalore, India.
- [22] Meinard Müller, Andreas Arzt, Stefan Balke, Matthias Dorfer, and Gerhard Widmer. 2018. Cross-modal music retrieval and applications: An overview of key methodologies. *IEEE Signal Processing Magazine* 36, 1 (2018), 52–62.
- [23] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. <https://doi.org/10.31115/v1-D14-1162>
- [24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*. PMLR, Online, 8748–8763.
- [25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Neural Information Processing Systems Workshop on Energy Efficient Machine Learning and Cognitive Computing*. Vancouver, BC, Canada.
- [26] Markus Schedl, Emilia Gómez, Julián Urbano, et al. 2014. Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval* 8, 2-3 (2014), 127–261.
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30.
- [28] Minz Won, Keunwoo Choi, and Xavier Serra. 2021. Semi-supervised music tagging transformer. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021*, Online.
- [29] Minz Won, Andrés Ferraro, Dmitry Bogdanov, and Xavier Serra. 2020. Evaluation of CNN-based Automatic Music Tagging Models. *Proceedings of Sound and Music Computing (SMC)* abs/2006.00751 (2020).
- [30] Minz Won, Justin Salamon, Nicholas J. Bryan, Gautham J. Mysore, and Xavier Serra. 2021. Emotion Embedding Spaces for Matching Music to Stories. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021*. Online.