# Are we describing the same sound? An analysis of word embedding spaces of expressive piano performance

Silvan David Peter
silvan.peter@jku.at
Johannes Kepler University
Linz, Austria

Shreyan Chowdhury
shreyan.chowdhury@jku.at
Johannes Kepler University
Linz, Austria

Carlos Eduardo Cancino-Chacón
carlos_eduardo.cancino_chacon@jku.at
Johannes Kepler University
Linz, Austria

Gerhard Widmer
gerhard.widmer@jku.at
Johannes Kepler University
Linz, Austria

## ABSTRACT

Semantic embeddings play a crucial role in natural language-based information retrieval. Embedding models represent words and contexts as vectors whose spatial configuration is derived from the distribution of words in large text corpora. While such representations are generally very powerful, they might fail to account for fine-grained domain-specific nuances. In this article, we investigate this uncertainty for the domain of characterizations of expressive piano performance. Using a music research dataset of free text performance characterizations and a follow-up study sorting the annotations into clusters, we derive a ground truth for a domain-specific semantic similarity structure. We test five embedding models and their similarity structure for correspondence with the ground truth. We further assess the effects of contextualizing prompts, hubness reduction, cross-modal similarity, and k-means clustering. The quality of embedding models shows great variability with respect to this task; more general models perform better than domain-adapted ones and the best model configurations reach human-level agreement.

## CCS CONCEPTS

• **Information systems** → **Test collections**; **Similarity measures**; **Top-k retrieval in databases**; • **Applied computing** → **Performing arts**.

## KEYWORDS

Semantic Similarity, Embeddings, Evaluation, Music Performance

**ACM Reference Format:**
Silvan David Peter, Shreyan Chowdhury, Carlos Eduardo Cancino-Chacón, and Gerhard Widmer. 2023. Are we describing the same sound? An analysis of word embedding spaces of expressive piano performance. In *Forum for Information Retrieval Evaluation (FIRE 2023), December 15–18, 2023, Panjim, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3632754.3632759



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

ACM ISBN 979-8-4007-1632-4/23/12.
https://doi.org/10.1145/3632754.3632759

## 1 INTRODUCTION

Semantic embeddings are a foundational concept in Natural Language Processing (NLP). NLP embedding models map words and their contexts to high-dimensional vector spaces while encoding as much of the semantic information as possible. Computers process such numerical data more readily than text, and vector spaces enable simple yet powerful similarity representations. Semantic embeddings enable a wide a variety of downstream tasks, from retrieval to classification to context-enhanced few shot learning. State-of-the-art (SOTA) embedding models are trained on vast datasets of natural language covering many domains and disciplines, underpinned by a general distributional hypothesis — that words with similar meanings occur in similar contexts.

This hypothesis and its corresponding inductive bias lead to one of the major open challenges related to semantic embeddings: whether context dependence, polysemy, and fine-grained domain-specific idiosyncracies are adequately represented by general purpose semantic embeddings. Specialized domains of language use such as talk of specific arts might exhibit incommensurable associations and similarites, e.g., a bright piano sound evokes different meanings than a bright student. While specific domains and their possibly nuanced differentiations do occur in large datasets, they only occur in their specific context, where the distributional hypothesis runs counter to different encoding. In other words, if certain emotionally dissimilar adjectives (happy, sad) only occur in similar contexts, they end up close in the embedding space despite opposing meanings.

In this article, we address the question whether general semantic embeddings can recover similarity relations in a specific domain of language use. In particular, we are interested in adjectival spaces used in the characterization of expressive performance of Western classical solo piano music. The underlying motivation is the possibility of expressivity- or emotion-based music retrieval using intuitive verbal queries, which would be a valuable and sought-after service in the digital music world.

Characterizations of expressive performance get at the finest details of performance technique, expression, timbre, emotions, metaphors, and associations. Note that it's crucially not the piece that is being described by such characterizations, but its expressive interpretation and rendition.

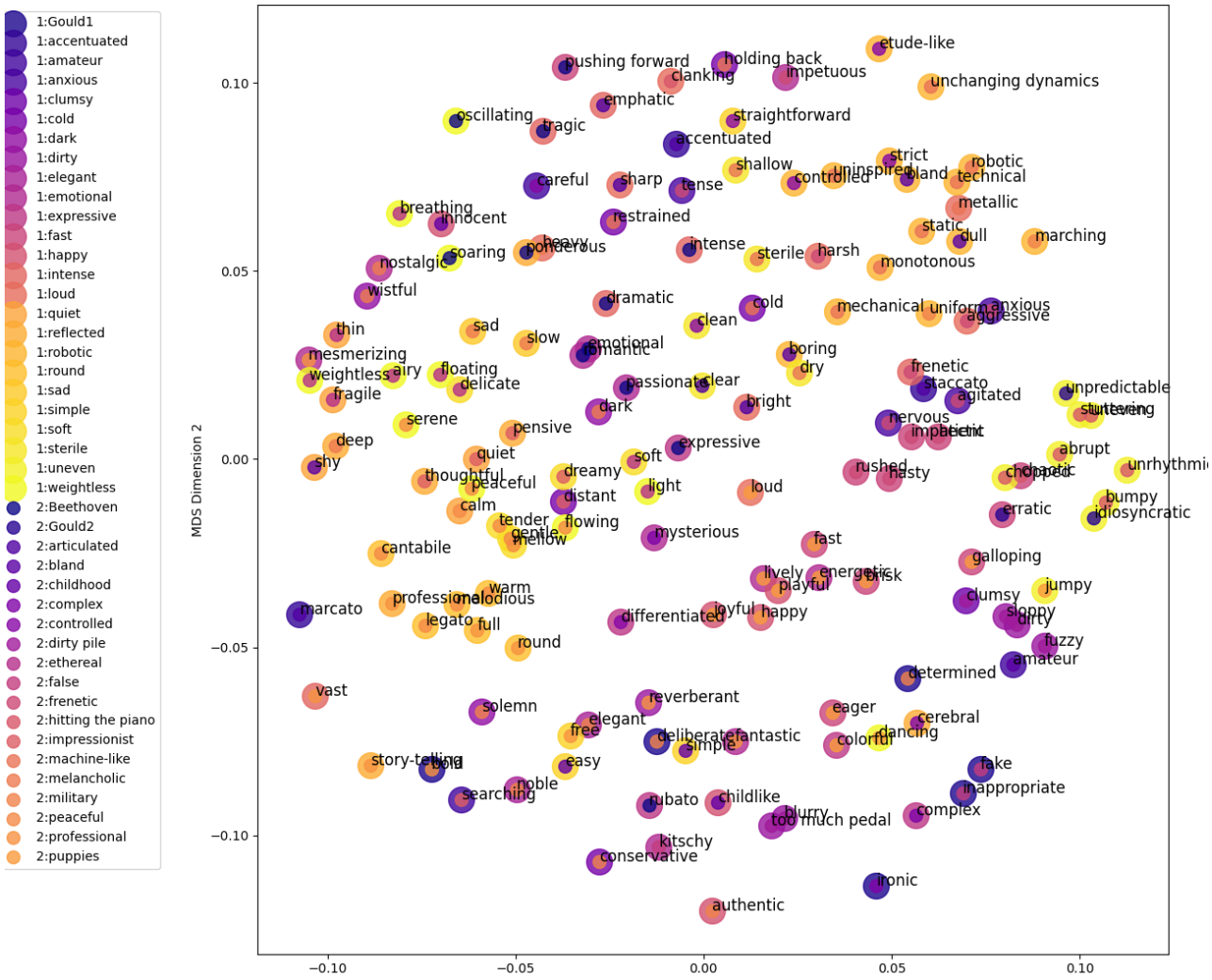


**Figure 1: Multidimensional Scaling (MDS) of the term data based the equally weighted pile group one, pile group two, and performance similarities. The legend on the left lists all piles of both groups, first the group number, then the names the musicians assigned them. Each term of the 150 in our ground truth data is shown in the scatter plot to the right and colored by the two piles it was sorted into, one for group one (large dots), one for group two (small dots). The musicians did not rate any similarities between piles, the color progressions for the piles do not encode closeness.**

Characterizations of expressive performance are both highly specific as well as very important to domain experts such as performers, teachers, and committed listeners. In fact, a crucial skill for aspiring performers consists in developing a sensibility as well as a language for performance nuances. Likewise, discriminating classical music lovers are highly sensitive to interpretation differences and can be very articulate in describing aspects of a performance that they don't like.

We use a dataset of terms used for the characterization of expressive performance in a large scale listening study. The dataset is annotated with similarity clusters of 150 terms created by two groups of domain experts in Western classical music performance. This data gives us an ecologically valid adjectival space of domain specific terms along with expert-annotated similarity annotations — an ideal experimental reference for general embedding spaces.

In this article, we take this reference similarity space of 150 terms related to the domain of expressive performance characterization, and compare it with embeddings for these terms derived from five embedding models by means of precision at k metrics. Furthermore, we present experiments investigating several factors affecting the embedding spaces: adding context to terms, reducing hubness in the embedding spaces, comparing audio to text embeddings, and a comparison of different clusterings in the embedding spaces.

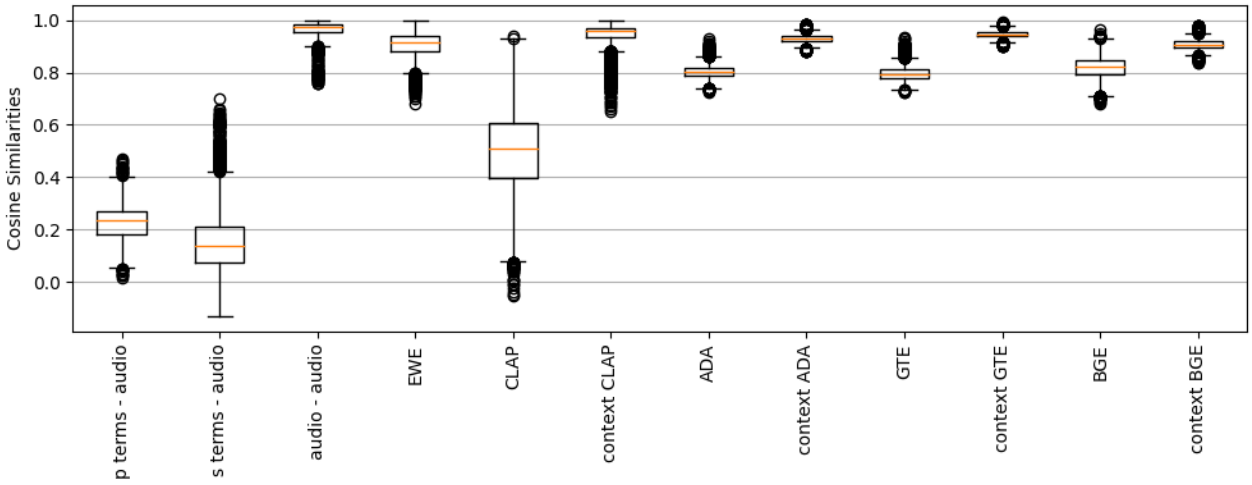


**Figure 2: Box plots of distributions of pairwise similarities in various embedding spaces. The main embedding models tested are labelled as EWE, CLAP, ADA, GTE, and BGE [1, 2, 13, 21, 25]. The three leftmost distributions relate to cross-modal audio and text embeddings as discussed in Section 4.2 and the distributions labeled with "context" are addressed in Section 4.4.**

## 2 RELATED WORK

The literature on semantic embeddings is vast and multifaceted, with many tasks and benchmarks making use of suitable language representations. For a recent overview and online benchmarking results, we refer to the Massive Text Embedding Benchmark (MTEB) [23, 24].

In our work, we investigate five models which we introduce in the following. These models cannot represent the whole of the state of the art, however, we do think they represent interesting models for our purpose. We use three models among the top performers in the MTEB. First, the general text embeddings (GTE) model "gte-large" developed by the Alibaba DAMO Academy [21]. This model is trained contrastively in both an unsupervised and a supervised fine-tuning fashion and as of September 2023 leads the MTEB leaderboard for semantic text similarity (STS) tasks. Second, we use the BAAI General Embeddings (BGE) model "bge-large-en" [2]. This model currently tops the overview MTEB leaderboard with minimal background information on the type of model and training available. Thirdly, we use OpenAI's general purpose embedding model "text-embedding-ada-002" offered at their API since December 2022 [7, 25]. Not many architectural details about this model are known, however, it performs in the top 20 models both in the overview as well as for STS tasks as of September 2023. Furthermore, it's likely one of the most widely used models in commercial applications.

We extend the model list with two specialized models: a pure word embedding model for emotion-enriched word embeddings (EWE) [1], trained to mitigate the inductive bias that emotion term embeddings are liable to be influenced by, and Microsoft's cross-modal text-audio embedding model (CLAP) [13], trained to embed both text and audio excerpts in the same space for cross-modal retrieval. We assume these models to be better suited to the language in the domains of audio and emotion description, respectively, both of which overlap with expressive performance characterization.

The last year has seen several publications investigating retrieval of perceptual structure information from large language models (LLM), and in particular using the model chatGPT. Most closely related to our proposal are ratings of timbral similarity [32], music similarity [17] , and general sensory judgment dimensions [22], all of which find evidence for the recovery of human annotations by chatGPT, albeit not at human level.

Our reference data is based on free text descriptions of expressive piano performance. People involved with music take pleasure in talking about music. This is no different for expressive performance of Western classical solo piano. In doing so they develop a rich vocabulary that relates to different aspects of performance such as evaluative/axiological terminology, emotions, metaphors, playing technique, or timbre descriptors. In music research, these aspects are often addressed separately, and with a generally reductionist approach. That is, researchers are interested in the identification of underlying factos, perceptual categories, and their relations to acoustic features [20, 26]. This is somewhat opposed to our approach, where no reduction of the semantic space is pursued. Nevertheless, we want to briefly outline two relevant areas of inquiry: emotions and timbre.

There exists a substantial literature investigating the emotional language related to music [11, 18]. Among prominent models are categorical models such as the Geneva Emotional Music Scale (GEMS) [34] and the dimensional valence and arousal model [27]. Crucial questions relate to the question whether musical emotions are perceived or induced [33] or the paradoxical enjoyment of negative emotions such a sadness [12]. Metaphors are another common linguistic device in the characterization of music. We refer to Schaerlaken et al. [30] who identified five main factor in their analysis of

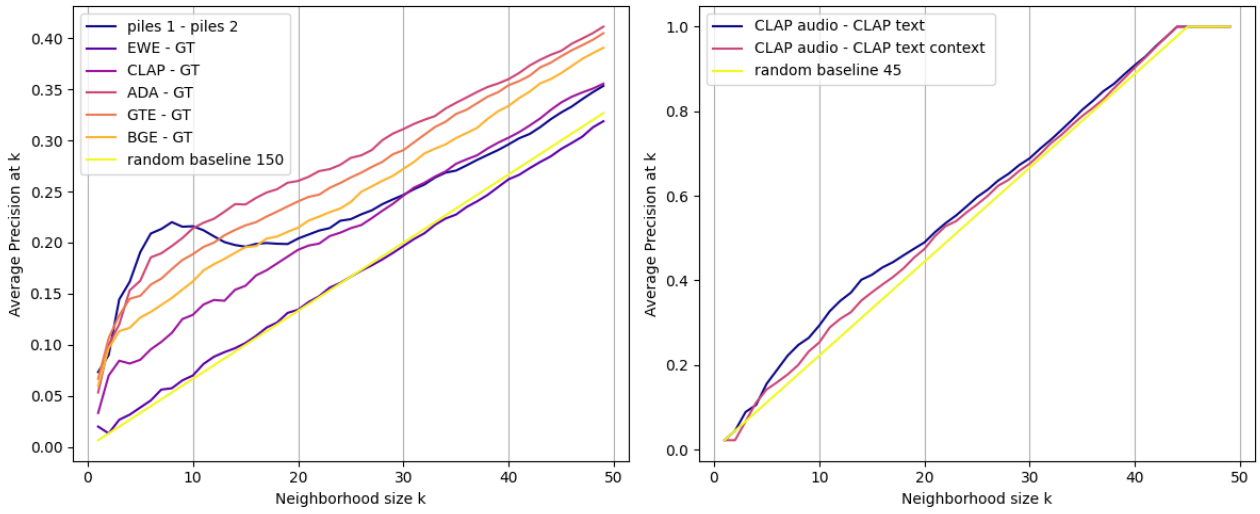


**Figure 3: Left plot: aP@k for k $\in \{1, ..., 49\}$ for several embeddings models against the ground truth similarities. Right plot: aP@k of 45 performance embeddings represented as CLAP audio embeddings and as mean CLAP text embeddings of terms (with and without context prompts).**

metaphorical attributed (GEMMES) and connected their perception to the GEMS [29].

Timbre is crucial topic in performance research and music psychology [14, 19, 28]. Timbre is increasingly conceptualized as controlled expressive performance attribute, i.e., as something that performers can influence with playing techniques and gestures. Most relevant to our work are several piano timbre description experiments by Bernays et al. [3–6] which brought five to eleven categorical terms to the fore. For our experiments, we rely on as many terms with associated similarity annotations as possible without reduction to principal factors or dimensions, which we find in the con espressione dataset and its pile sorting extension, detailed in Section 3.1.

## 3 METHODS

For our experiments we require an embedding space of words (for simplicity: adjectives) with associated ground truth similarity ratings as a type of ground truth reference data. We also need several SOTA embedding models and the metrics by which we can meaningfully compare the resulting similarity structures. In this section, we introduce these components.

### 3.1 Con Espressione Data

The Con Espressione Dataset (CED) collected descriptions of piano performances through an online questionnaire [9, 10]. Participants listened to 45 different performances by famous pianists of nine excerpts of Western classical piano pieces. The participants were shown the prompt: 'Please think of words (if possible, adjectives) that best describe the character of each performance to you.' and a text field allowed for free text answers (as many words as they liked, in a language of their choice). They were further instructed to concentrate on the performance aspects and not on the piece of music itself. The CED characterizations contains 3,166 terms, of which 1,415 are unique. Consequently, the CED consists of very loosely structured text data for which relational semantic ground truth, i.e. which answers refer to the same aspect or idea, is largely missing.

To mitigate this, a follow-up experiment was designed [8]. In two separate sessions, groups of professional musicians/musicologists sorted the 150 most often occurring terms in the CED into piles. These piles should cluster terms that describe a common expressive character. The type of similarity and number of piles were left open. The two groups of four professional musicians each sorted the terms into 25 and 19 piles, respectively, and gave a name to each pile, in the form of an adjective that best summarizes the common meaning of the words associated with the pile. [1]

### 3.2 Embeddings and Similarities

We use the same 150 terms as our adjectival space. To derive ground truth similarities between the terms we use both term co-occurrence in piles (from both groups) and co-occurrence for performance descriptions. For each of the pile groups, we create a similarity matrix where pairwise similarities of terms within a pile are set to one, outside the pile to zero. Term similarities are also computed based on the CED directly, where two different terms that occur in the characterization of the same performance are assumed more similar than if they are used for different performance. We again compute a similarity matrix where co-occurrence of two terms in the same performance description sets the similarity to one, else to zero. We weigh each source of similarity annotations the same, that is, the similarities from pile group one, pile group two, and performance description are summed up and finally normalized.

[1] An interactive interface for the exploration of piles, terms, and performances is available at: https://cpjku.github.io/expressivity/

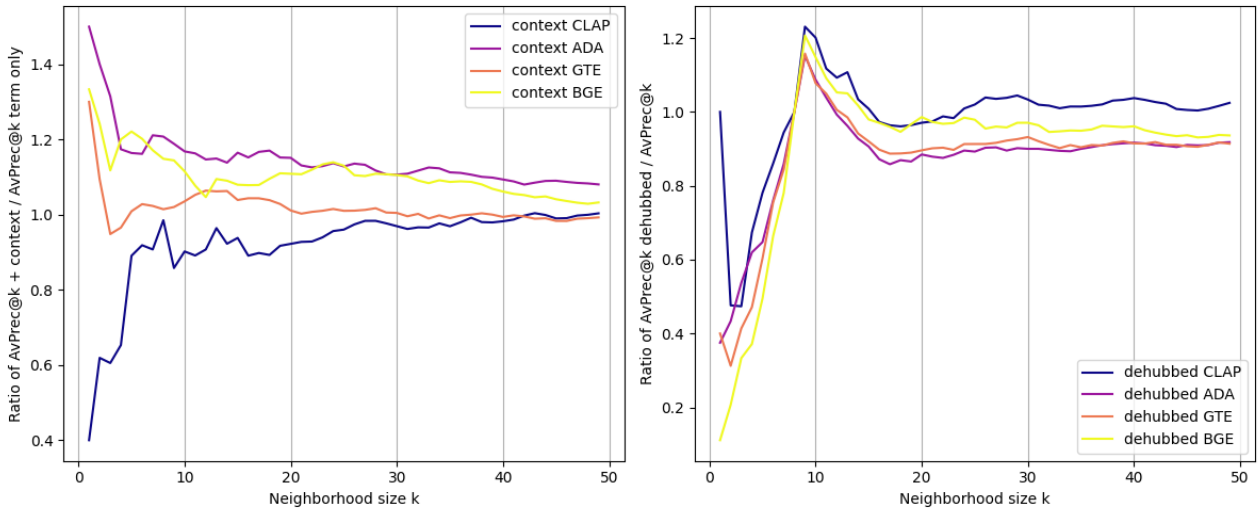


**Figure 4: Left plot: relative change in aP@k brought about by the inclusion of contextualizing prompts. Right plot: relative change in aP@k due to hubness reduction at neighborhoods of size eight.**

Figure 1 illustrates the resulting similarity structure. The legend on the left lists all piles of both groups (with the names the musicians gave them) the scatter plot on the right shows a low-dimensional approximation of the term similarities.

We compare our ground truth similarities with similarities for the same adjectival space, i.e., the same 150 terms, embedded using five different embedding models: ADA (embedding dimensionality 1536), CLAP (1024), EWE (300), GTE (1024), and BGE (1024), detailed in Section 2.

### 3.3 Metrics

We assess the similarities using several metrics. All similarities are cosine similarities in the embedding spaces:

$$CosineSimilarity(x, y) = \frac{x \cdot y}{|x||y|} \tag{1}$$

for $x$ and $y$ term embeddings in the same space.

Our main evaluation metric is the average precision at k (aP@k) for k nearest neighbors. For each term x in our adjectival space S, we compute two neighborhoods of size k, one according to our ground truth embedding space U (knn(x,U)), and one according to a test embedding space V (knn(x,V)). We then compute the precision of retrieval of the U-neighborhood by the V-neighborhood:

$$aP@k(U, V) = \frac{1}{|S|} \sum_{x \in U} \frac{|knn(x, U) \cap knn(x, V)|}{k} \tag{2}$$

where knn(x,U) denotes the set of k nearest neighbors of x according to the embedding space U. We choose aP@k over the more common Spearman correlation of embedding similarities for STS tasks [23], for its capacity to naturally differentiate each level of neighborhood size k and its rank-agnostic behavior within the neighborhoods, which corresponds to the ground truth pile sortings.

In Section 4.5, we compare the agreements between the pile groups (ground truth sets of clusters) with k-means clusterings in the embedding spaces. For this purpose, we compute the overlap coefficient between sets of clusters according to different similarities. In particular, we compute the average maximal overlap between two different sets of clusters $\mathbb{C}_1$ and $\mathbb{C}_2$ as:

$$AvMaxOverlap(C_1, C_2) = \frac{1}{|\mathbb{C}_1|} \sum_{C_1 \in \mathbb{C}_1} \max_{C_2 \in \mathbb{C}_2} \frac{|C_1 \cap C_2|}{min(|C_1|, |C_2|)} \tag{3}$$

Note that this metric is not symmetrical in $\mathbb{C}_1$ and $\mathbb{C}_2$.

## 4 EXPERIMENTS AND RESULTS

This section presents experiments addressing structural correspondences between similarity spaces. After a main experiment in Section 4.1, we investigate more fine-grained issues regarding the correspondence of text and audio embeddings in Section 4.2 the effects of hubness in the embedding spaces in Section 4.3, the effects of contextualized terms in Section 4.4, and the overlap in clusterings in Section 4.5.

### 4.1 Similarity Structure Correspondence

To what extent do the models' embedding spaces correspond to the similarity relations derived from expert annotations? We answer this question using aP@k values. However, we first make a few general observations about the distributions of similarities without reference to structural considerations.

Figure 2 shows box plots of the distributions of pairwise similarities in several spaces under scrutiny. Note that for all language models (labelled as EWE, CLAP, ADA, GTE, and BGE) the embeddings are very similar, with values rarely falling below 0.7 and mean values above 0.8. This illustrates the inductive bias of the embeddings as these terms are likely to occur in similar and sometimes very particular domains.

Figure 3 (left) shows the performance of each embedding in terms of its approximation of the neighborhood structure of a reference

embedding. For neighborhood sizes k between one and 49, the aP@k of each embedding space is computed against our ground truth data. The models are are clearly ranked for all k, from top to bottom: ADA, GTE, BGE, CLAP, and EWE. A yellow diagonal denotes a random baseline. With the exception of EWE are all models significantly better than random. The upper bound of the 95% confidence interval (not shown) of the random baseline is approximately 0.02 above the baseline itself at k=1 and gets closer to the baseline as k increases.

A blue line indicates the only values that are not compared against the ground truth. Instead it compares similarities only based on pile group 1 (co-occuring terms in a pile have similarity one, else zero) with similarities only based on pile group 2. This is added as an indication of where human agreement about similarity might fall, albeit with some caveats: The values are only really meaningful for k in the approximate size of piles (roughly 5-10 terms/pile). For smaller k, the piles do not encode greater or lesser similarity of terms within a pile, all values are one and an arbitrary ordering was created by addition of some minimal noise. For larger k, the piles do not encode greater or lesser similarity for terms beyond the pile, all values are zero and again minimal noise was added for an arbitrary ordering. It's also not possible to compare these similarities against the ground truth since they were part of the creation of the ground truth which distorts the result unfairly. The blue line segment for k = 5 - 10 does give an indication that none of the models reach the agreement of two groups of expert annotators.

## 4.2 Audio vs. Text Embeddings

The CLAP model is trained to minimize distances between corresponding text and audio embeddings which makes it possible to compare against audio embeddings of the 45 performances that are characterized by the 150 terms in the con espressione data. The three leftmost distributions in Figure 2 are related to this approach: "audio - audio" denotes the pairwise similarities between audio embeddings, "s terms - audio" shows cross-modal similarities between 150 individual terms and 45 performances, and "p terms - audio" shows cross-modal similarities between 45 performances averaged from their corresponding term embeddings (the terms used to describe the performance in the CED) and 45 performance recording embeddings. Figure 3 (right) shows the latter in an aP@k plot; the 45 performances are embedded in both in the text spaces of two different text models (see Section 4.4 for a discussion on the "context" model) and in the audio space. The values are notably lower, for several k the values are not significantly better than the random baseline, indicating that either the audio or the text space do not have the granularity to represent these minute differences. We conjecture that the audio model is the more likely source of misalignment. After all, all performance recordings can reasonably be classified as "classical solo piano" which is closer to the level of precision to be expected from an unspecific audio representation model trained on a variety of (non-)musical audio [13].

## 4.3 The Effect of Hubness

The existence of hubs has repeatedly been found a source of distorted similarity structures, especially in high-dimensional spaces [15].

| Model | nbhd | Skewness | | Robinhood | |
|---|---|---|---|---|---|
| ADA | 4 | 0.99 | 0.88 | 0.25 | 0.21 |
| CLAP | 4 | 1.09 | 0.52 | 0.22 | 0.19 |
| GTE | 4 | 1.11 | 0.91 | 0.24 | 0.20 |
| BGE | 4 | 1.76 | 1.14 | 0.32 | 0.29 |
| RB | 4 | 0.54 | 0.40 | 0.17 | 0.13 |
| ADA | 8 | 1.12 | 0.58 | 0.25 | 0.21 |
| CLAP | 8 | 1.19 | 0.54 | 0.25 | 0.23 |
| GTE | 8 | 1.13 | 0.44 | 0.24 | 0.21 |
| BGE | 8 | 1.97 | 1.63 | 0.39 | 0.36 |
| RB | 8 | 0.37 | 0.17 | 0.13 | 0.11 |
| ADA | 16 | 0.81 | 0.32 | 0.26 | 0.24 |
| CLAP | 16 | 1.21 | 0.70 | 0.25 | 0.23 |
| GTE | 16 | 0.67 | 0.24 | 0.25 | 0.24 |
| BGE | 16 | 1.83 | 1.61 | 0.43 | 0.42 |
| RB | 16 | 0.25 | 0.04 | 0.10 | 0.08 |

**Table 1: Results of hubness measurement and reduction. Values for each model (RB random baseline) are presented for three different neighborhood sizes (nbhd). All skewness and robinhood values are doubled, left original, right after hubness reduction.**

Hubs are points that appear too often in k nearest neighborhoods of other points and as such are liable to influence the aP@k. In this experiment we address the influence of hubs for four models.

We first measure hubness as both skewness of the k-occurrence histogram (higher skewness indicates more hubness) and as robinhood index (which indicates the percentage of slots in nearest neighbor lists would need to be redistributed for equal distribution). We carry this computation out for three neighborhood sizes (4,8, and 16) and compute hubness reduction using an approximate mutual proximity method [31]. For algorithmic details regarding hubness measurement and reduction we refer the reader to Feldbauer et al. [16].

Table 1 shows the results of hubness reduction. For each metric (skewness, robinhood) we note two values for each setting: before (left) and after (right) hubness reduction. Several models show significant hubness (skewness > 1.0, robinhood > 0.25), with BGE being a negative outlier, and they almost universally benefit from reduction.

What does this mean for similarity structure recovery against the ground truth? Figure 4 (right) shows the relative change in aP@k for 4 models after hubness reduction at neighborhood size eight. This neighborhood was chosen for being crucial against ground truth based on piles, which on average have approximately this size. Notably, hubness reduction leads to a decrease for k less than the set neighborhood. For values around eight, hubness reduction universally leads to an increase in performance of about 20%, enough to boost the highest performing models an the league of the expert agreement baseline in this crucial area (see Figure 3 on the left, see section 4.1).

## 4.4 The Effect of Context

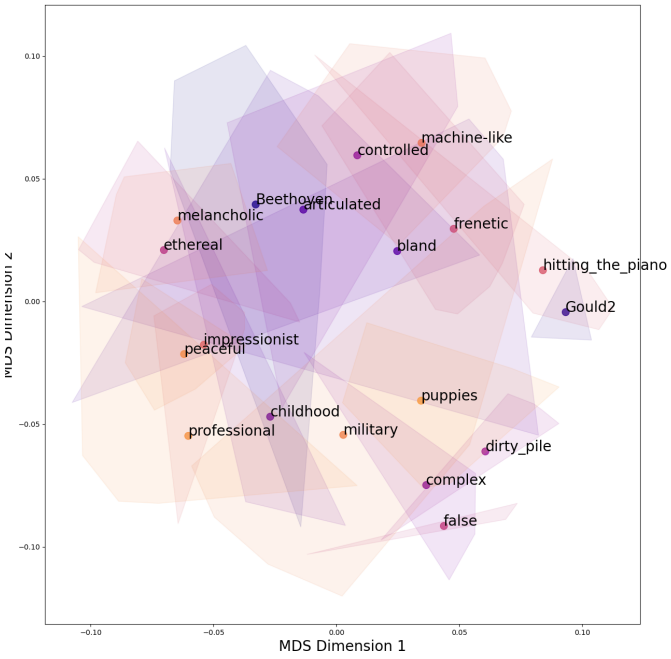


**Figure 5: Visualization of the convex hull of terms of each pile as embedded in the ground truth data. MDS dimension reduction for illustrative purposes only, 8+ dimensions are required to represent the space with minimal loss of information (<10% reduction in aP@k against original data). The pile centers are shown as average term embedding positions and annotated with the pile names given by the musicians.**

SOTA text embedding models are often capable of encoding term contexts into their representations, i.e., terms can be augmented with suitable prompts to specify the context of use. Such contextualizing information ideally enables the model to learn domain-specific similarities and relations that might not be apparent in other situations, such as metaphorical usage of terms related to movement, weight, and flow, or words borrowed from other sensory modalities like sweet, rough, and warm, which are common in our ground truth data. In this experiment, we augment the 150 terms with a common context prompt. In the original Con Espressione game, listeners were asked to 'please think of words (if possible, adjectives) that best describe the character of each performance to you.'. We translate this to the prompt: 'I listen to a solo piano performance of a classical piece of music and I'd describe the character of the performance as TERM' and recompute all embeddings for four test models.

Figure 2 shows the distributions of pairwise similarities for context prompts. For all tested models, the prompts led to higher pairwise similarities. To test the similarity structure of these context embeddings, we again compute the relative change in aP@k for four models after adding contexts. Figure 4 (left) shows this relative change, i.e., the aP@k of the embeddings with context divided by the aP@k of those without. Not all models react positively context information, GTE and CLAP stay largely the same or get worse.

| Ref | P1 in P2: | 0.62 | P2 in P1: | 0.65 |
|---|---|---|---|---|
| **Model** | **Overlap v P1** | | **Overlap v P2** | |
| ADA | 0.55 | 0.58 | 0.48 | 0.43 |
| CLAP | 0.52 | 0.51 | 0.47 | 0.46 |
| GTE | 0.59 | 0.59 | 0.53 | 0.52 |
| BGE | 0.51 | 0.53 | 0.53 | 0.48 |
| GT | 0.66 | 0.70 | 0.66 | 0.70 |
| RB | 0.39 | 0.41 | 0.40 | 0.39 |

**Table 2: Overlap and distance ratios.**

The other two models see performance increases of 20 % and more, which in the case of ADA pushes the model higher than the expert reference (see Figure 3, see section 4.1).

## 4.5 Clustering and Piles

The aP@k compares neighborhoods of the same size, however, the experts' groups of piles are not homogeneous in size and number. On the other hand, the piles do provide a complete clustering of the adjectival space which can be compared against automatic clusterings (see Figure 5 for an illustration of the clustering provided by pile group two in a the ground truth space). In this last experiment, we compute k-means clusterings in several embedding spaces and compare them against the two groups of piles by means of average maximal overlap coefficients.

Table 2 shows the results. All k-means clusterings are computed with k=22, the average of the pile group sizes ($|P1| = 25, |P2| = 19$). Average maximal overlap coefficients are not symmetrical, hence we report two values per setup. The top row reports two values for reference: the average maximal overlap of pile group one with pile group two ("P1 in P2"), and vice versa ("P2 in P1"). K-means clusterings are computed for six models, four embedding models, the ground truth model, and a random baseline consisting of a 100-dimensional Gaussian. The two columns below "Overlap v P1" average maximal overlap coefficients for group one in k-means clusters (left) and k-means clusters in group one (right). The next two columns marked "Overlap v P2" report the same values for group two.

Note that smaller sets of clusters generally reach a higher average maximal overlap due to larger clusters (P2 in P1 > P1 in P2, $|P1| = 25$, $|P2| = 19$). None of the embedding models reach the agreement between the two groups of piles, however, they clearly outperform the random baseline. The ground truth reaches higher overlap values with both groups than in between the groups, which is to be expected, as it was derived from the performance annotations and the two groups of piles.

# 5 DISCUSSION AND CONCLUSION

Specialized datasets created by domain experts like the CED and its sorted pile groups usually serve a primary reductionist research purpose: the identification and description of dimensions, categories, and relations in perceptual-linguistic spaces. However, they also allow for a quantitative glimpse into the similarity structures of these spaces: similarity structures which are hypothesized to

be recovered by SOTA term embedding models. In a series of experiments, we address this hypothesis for a presumably highly specialized type of adjectival space, that of characterizations of expressive performance of Western classical solo piano works.

Our results show that domain-specific semantic similarity structures are indeed represented in the embedding spaces — to a degree. The tested models span the full range from near the random baseline to near human agreement. General-purpose models perform better than domain-adapted ones, running counter to our initial assumption.

In our experiments, cross-modal audio embeddings of the performance recordings fail to exhibit the same similarity structure as text embeddings. Hubness reduction helps the similarity correspondence universally, albeit only for a specific and small segment of neighborhoods. The inclusion of contextualizing prompts affects the models differently, with the best models receiving a clear boost in quality. Overlap statistics show that all embedding space clusters show less correspondence with the expert sortings than those show among themselves.

The groups of piles are used as a reference for correspondence in similarity structures throughout as they represent a rough estimate of inter-rater agreement to be expected. They are however the largest source of uncertainty. We do not know how likely similar groups of piles are or have other means of assessing of inter-rater agreement. Neither do we have direct similarity ratings or know whether performance expressivity-specific similarities are indeed notably different from general similarity. Research into dimensional and categorical structure on music perception as discussed in section 2 may ground the pile group similarities, however, for the number of terms used or even the free-text CED annotations more research is required to illuminate the robustness of this data.

To conclude, general state-of-the-art text embedding models can show correspondence with expert annotated perceptual-linguistic similarities that reach the experts' inter-rater agreement while other — even plausibly better suited domain models — fail at this task. Future research includes the investigation of the robustness of the annotation data as well as the extension of this approach to other domains where fine-grained and possibly idiosyncratic adjectival spaces are used.

## 6 REPRODUCIBILITY

Our data and code is available at:
https://github.com/CPJKU/performance_embeddings_fire23

## ACKNOWLEDGMENTS

This work is supported by the European Research Council (ERC) under the EU's Horizon 2020 research & innovation programme, grant agreement No. 101019375 ("Whither Music?"), and the Federal State of Upper Austria (LIT AI Lab).

## REFERENCES

[1] 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.
[2] BAAI. 2023. BGE repository. https://github.com/FlagOpen/FlagEmbedding
[3] Michel Bernays and Caroline Traube. 2010. Expression of piano timbre: gestural control, perception and verbalization. In *Proceedings of CIM09: The 5th Conference on Interdisciplinary Musicology*.
[4] Michel Bernays and Caroline Traube. 2011. Verbal expression of piano timbre: Multidimensional semantic space of adjectival descriptors. In *Proceedings of the international symposium on performance science (ISPS2011)*. European Association of Conservatoires (AEC) Utrecht, Netherlands, 299–304.
[5] Michel Bernays and Caroline Traube. 2013. Expressive production of piano timbre: touch and playing techniques for timbre control in piano performance. In *Proceedings of the 10th Sound and Music Computing Conference (SMC2013)*. KTH Royal Institute of Technology Stockholm, Sweden, 341–346.
[6] Michel Bernays and Caroline Traube. 2014. Investigating pianists' individuality in the performance of five timbral nuances through patterns of articulation, touch, dynamics, and pedaling. *Frontiers in Psychology* 5 (2014), 157.
[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[8] Carlos Cancino-Chacón, Silvan Peter, Shreyan Chowdhury, Anna Aljanaki, and Gerhard Widmer. 2021. Sorting Musical Expression: Characterization of Descriptions of Expressive Piano Performances. In *Extended Abstract in 16th International Conference on Music Perception and Cognition ICMPC 2021) and 11th Triennial Conference of ESCOM (ICMPC-ESCOM 2021)*.
[9] Carlos Cancino-Chacón, Silvan Peter, Shreyan Chowdhury, Anna Aljanaki, and Gerhard Widmer. 2023. Con Espressione Dataset. https://zenodo.org/record/3968828
[10] 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.
[11] Tuomas Eerola and Jonna K Vuoskoski. 2012. A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception: An Interdisciplinary Journal* 30, 3 (2012), 307–340.
[12] Tuomas Eerola, Jonna K Vuoskoski, Henna-Riikka Peltola, Vesa Putkinen, and Katharina Schäfer. 2018. An integrative review of the enjoyment of sadness associated with music. *Physics of Life Reviews* 25 (2018), 100–121.
[13] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
[14] Philippe Esling, Axel Chemla-Romeu-Santos, and Adrien Bitton. 2018. Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics. arXiv:1805.08501 [cs.SD]
[15] Roman Feldbauer and Arthur Flexer. 2019. A comprehensive empirical comparison of hubness reduction in high-dimensional spaces. *Knowledge and Information Systems* 59, 1 (2019), 137–166.
[16] Roman Feldbauer, Maximilian Leodolter, Claudia Plant, and Arthur Flexer. 2018. Fast approximate hubness reduction for large high-dimensional data. In *2018 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, 358–367.
[17] Arthur Flexer. 2023. Can ChatGPT be useful for distant reading of music similarity?. In *Proceedings of the 2nd Workshop on Human-Centric Music Information Retrieval (HCMIR), 2023*.
[18] Alf Gabrielsson. 2003. Music performance research at the millennium. *Psychology of music* 31, 3 (2003), 221–272.
[19] Carol L Krumhansl. 1989. Why is musical timbre so hard to understand. *Structure and perception of electroacoustic sound and music* 9 (1989), 43–53.
[20] Alexander Lerch, Claire Arthur, Ashis Pati, and Siddharth Gururani. 2020. An Interdisciplinary Review of Music Performance Analysis. *Transactions of the International Society for Music Information Retrieval* (Nov 2020). https://doi.org/10.5334/tismir.53
[21] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-stage Contrastive Learning. arXiv:2308.03281 [cs.CL]
[22] Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. 2023. Large language models predict human sensory judgments across six modalities. arXiv:2302.01308 [cs.CL]
[23] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive Text Embedding Benchmark. (2022). arXiv:2210.07316 [cs.CL]
[24] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB Leaderboard. https://huggingface.co/spaces/mteb/leaderboard
[25] OpenAI. 2022. New and improved embedding model. https://openai.com/blog/new-and-improved-embedding-model
[26] Caroline Palmer. 1997. Music performance. *Annual review of psychology* 48, 1 (1997), 115–138.
[27] James A Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.
[28] Charalampos Saitis, Stefan Weinzierl, Katharina von Kriegstein, Sølvi Ystad, and Christine Cuskley. 2020. Timbre semantics through the lens of crossmodal correspondences: A new way of asking old questions. *Acoustical Science and Technology* 41, 1 (2020), 365–368.

[29] Simon Schaerlaeken, Donald Glowinski, and Didier Grandjean. 2022. Linking musical metaphors and emotions evoked by the sound of classical music. *Psychology of Music* 50, 1 (2022), 245–264.

[30] Simon Schaerlaeken, Donald Glowinski, Marc-André Rappaz, and Didier Grandjean. 2019. "Hearing music as...": Metaphors evoked by the sound of classical music. *Psychomusicology: Music, Mind, and Brain* 29, 2-3 (2019), 100.

[31] Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer. 2012. Local and global scaling reduce hubs in space. *Journal of Machine Learning Research* 13, 10 (2012).

[32] Kai Siedenburg and Charalampos Saitis. 2023. The language of sounds unheard: Exploring musical timbre semantics of large language models. arXiv:2304.07830 [cs.CL]

[33] Yading Song, Simon Dixon, Marcus T Pearce, and Andrea R Halpern. 2016. Perceived and induced emotion responses to popular music: Categorical and dimensional models. *Music Perception: An Interdisciplinary Journal* 33, 4 (2016), 472–492.

[34] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. 2008. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion* 8, 4 (2008), 494.