Audio, Lyrics, Videoclips, Interactions? An Analysis of Uni- and Multi-modal Music Retrieval Systems in Terms of Accuracy and Beyond-accuracy Aspects

Marta Moscati[1,*], Gustavo Escobedo[1], Eduardo Hernandez Almanza[1], Jonas Peché[1] and Markus Schedl[1,2]

[1]Johannes Kepler University Linz, Linz, Austria
[2]Linz Institute of Technology, Linz, Austria

Abstract

Representations of the audio content of music tracks and of related data (such as lyrics, user-generated tags, or interaction data) are often leveraged in music retrieval and recommendation systems. It is therefore important to know how the choice of representation affects the results of similarity-based music retrieval systems. In this work, we address this question under several aspects. We analyze the accuracy, coverage, hubness, popularity bias, and robustness of retrieval systems based on different content modalities (text, audio, video) and on user–item interactions, and analyze the impact of corresponding features on multimodal retrieval systems. The paper gives useful insight into which modality to leverage depending on the aspects of retrieval results to prioritize and hence provides guidelines for practical real-world scenarios.

Keywords

Music similarity, Music information retrieval, Accuracy, Coverage, Popularity bias, Hubness, Robustness, Content features, Collaborative data, Evaluation study

1. Introduction

The way music listeners access music tracks is diverse. Some listeners prefer the use of video or music streaming platforms, while others prefer purchasing albums. This is reflective of the fact that, although the production of music is most naturally related to the audio signal, music producers also devote significant efforts in designing additional content of the music tracks, such as album covers or videoclips. Correspondingly, music listeners also select which music to listen to based on several modalities. This renders the way the similarity between music tracks is perceived intrinsically multimodal. Additionally, the amount of music available is vast and ever-increasing, which renders music retrieval systems essential for supporting listeners in selecting relevant music tracks.

In this work, we analyze the performance of different representations of music in the task of retrieving music tracks that are similar to a query track. We consider retrieval systems based on the audio signal, the lyrics, or the videoclips of the tracks, as well as on user–item interactions from music streaming platforms collected through the music website Last.fm.[1] Additionally, we include multimodal systems based on early- or late-fusion. We analyze the performance of retrieval systems in terms of *accuracy* and *beyond-accuracy* aspects. In particular, we measure the ability of retrieval systems to capture aspects that define music similarity in terms of music genres, as commonly done in music information retrieval (MIR) research [1]. Since genres are not mutually exclusive, to balance the skewness in the distribution of genres over tracks, we include definitions of relevance that are binary or continuous-valued, based on measures for the similarity of sets. We also include in our analysis catalog *coverage*, *popularity bias*, and

MuRS 2025: 3rd Music Recommender Systems Workshop, September 22nd, 2025, Prague, Czech Republic
*Corresponding author.
marta.moscati@jku.at (M. Moscati); gustavo.escobedo@jku.at (G. Escobedo); eduardohdz0697@hotmail.com (E. H. Almanza); j_peche@wargaming.net (J. Peché); markus.schedl@jku.at (M. Schedl)
0000-0002-5541-4919 (M. Moscati); 0000-0002-4360-6921 (G. Escobedo); 0009-0004-5683-8213 (E. H. Almanza); 0009-0001-8250-9008 (J. Peché); 0000-0003-1706-3406 (M. Schedl)

[1]https://www.last.fm

hubness, i.e., the tendency of retrieving a small number of the same tracks over and over again, since these have often been considered particularly relevant to MIR applications [2, 3, 4, 5, 6]. To analyze the robustness of modalities and the impact of individual modalities on multimodal systems, we quantify the *coherence* of the retrieval results within and across different modalities. This information gives insight into the amount of change of the retrieval results when replacing one (feature from one) modality with another, and hence helps in providing guidance for real-world scenarios where, e.g., one feature or one modality is not available. We create an interactive dashboard[2] to allow deeper explorations of the results of our analysis and provide the code to reproduce our experiments.[3]

The remainder of the paper is organized as follows: In Section 2 we discuss previous work related to ours, in particular regarding similarity-based music retrieval methods and their evaluation (Section 2.1), and regarding beyond-accuracy metrics in MIR domains (Section 2.2). In Section 3 we provide the mathematical formulation of the retrieval task, describe the methodology underlying the retrieval systems, and the metrics used to evaluate their performance in terms of accuracy- and beyond-accuracy aspects. In Section 4 we describe our experiment setup, namely the dataset, the features used for the retrieval systems, and the approach adopted to create the collaborative filtering (CF) representations. We report the results of our experiments and discuss our observations in Section 5. Finally, we discuss the limitations and possible extensions of our work in Section 6.

2. Related Work

In this section, we briefly present work on similarity-based music retrieval systems and in particular on the comparison of different features in MIR tasks, as well as on beyond-accuracy metrics in MIR.

2.1. Similarity-based music retrieval

Similarity-based music retrieval, i.e., the task of ranking the tracks of a music catalog based on the similarity to the query track [7], is the basis of many music delivery applications [7, 8, 9]. Standard techniques for similarity-based music retrieval rely on unsupervised approaches [10, 11, 12] or supervised approaches [8, 13] that use user-generated data such as tags as learning signals [14]. Recent work also leverages self-supervised [7, 15] and unsupervised learning based on contrastive losses [16, 17]. Su et al. [18] systematically evaluate the impact of the parameters of bag-of-frames representations of the audio signal on several MIR tasks, such as genre classification and pitched instrument recognition. More recently, Plachouras et al. [19] introduce a framework to evaluate representations of the audio signal on several MIR tasks and datasets, including robustness to perturbations in their analysis. Our analysis differs from previous studies in that we focus on the task of music retrieval, which has not been considered by Su et al. [18] nor Plachouras et al. [19], and which is closely connected to industry domains such as that of music recommendation. Furthermore, we consider a multimodal scenario and in addition to representations of the audio we also include in our analysis representations of the lyrics, of the videoclips, and of user–item interaction data. Finally, our evaluation extends to aspects beyond accuracy.

2.2. Beyond-Accuracy Evaluation

One of the areas of MIR in which beyond-accuracy aspects are gaining increasing attention is that of music recommendation. Music recommender systems (MRSs) [2, 3, 9] are one of the main applications of similarity-based music retrieval, and several works [20, 21, 22] highlight the importance of measuring aspects of the quality of recommendation that go beyond accuracy. Among those, catalog coverage [23], hubness [24, 5, 6], and popularity bias [4] are of particular relevance to MRSs. However, these aspects are typically neglected in other MIR tasks such as similarity-based music retrieval. The work at hand differs

[2]Dashboard: tinyurl.com/cmrs2024
[3]Code:https://github.com/hcai-mms/multimodal_mir

from previous work on beyond-accuracy evaluation of MIR systems. In fact, we analyze similarity-based music retrieval systems under aspects typically not jointly considered in their evaluation.

3. Methodology

In this section, we provide a mathematical definition of the retrieval systems (Section 3.1) and of their evaluation (Section 3.2).

3.1. Retrieval Systems

Given the catalog M, a music track $m \in M$ is represented by several *feature vectors* $\mathbf{f_k^{(m)}} \in \mathbb{R}^{n_k}$, where k is an index labelling the feature and n_k is the dimensionality of the corresponding vector. A *retrieval system* $\phi^{(\mathbf{f_k},\text{dis})}$ is defined by the combination of feature $\mathbf{f_k}$ and distance in feature space (dis). Given a *query track* $q \in M$, the retrieval system $\phi_N^{(\mathbf{f_k^{(q)}},\text{dis})}$ returns the N *retrieved tracks* $[r_1^{(q)}, \dots, r_N^{(q)}] \in M^N$ that have the lowest distance with q, breaking ties randomly. We refer to the audio signal, the lyrics, and the videoclip of the track as *modalities* that represent *content information*. We consider *unimodal* retrieval systems based on a feature $\mathbf{f_k}$ representing a single modality. We also include *multimodal* retrieval systems that simultaneously leverage one feature from the audio, one from the lyrics, and one from the videoclip modality using early- or late-fusion.[4] In addition to content representations, we consider three representations of user–item interaction data created using CF algorithms. Two are based on traditional recommendation algorithms, item k-nearest-neighbors (`ItemkNN`) and matrix factorization with truncated singular value decomposition (`SVD`). One is based on a well-established neural network (NN) architecture for recommendation, multinomial variational autoencoder (`MultVAE`), selected for its accuracy in the task of music recommendation [25]. For retrieval systems based on `ItemkNN`, we represent each track as the corresponding item vector in the user–item interaction matrix. For `SVD` we represent tracks with the embeddings multiplied by the square root of the singular values. `MultVAE` is usually trained to encode and reconstruct the user profiles. Since we are interested in the track representations, we use the same architecture to reconstruct the track profiles. Therefore, we train an instance of `MultVAE` on the transposed user–item interaction matrix and use the latent vectors of the tracks as features in the retrieval system. In initial experiments, we considered either inverted cosine or Tanimoto similarity as distances. Since all retrieval systems reached higher accuracy with cosine similarity, we restrict our discussion at hand to retrieval systems based on cosine.

3.2. Evaluation

We measure the *accuracy* of a retrieval system in terms of normalized discounted cumulative gain (NDCG) with gain based on the genres of the query q and the retrieved tracks $[r_1^{(q)}, \dots, r_N^{(q)}]$. Each track $q, r_j^{(q)} \in M$ is labeled with a subset G_m of the set of all genres, $G_m \subset G$. We consider the j-th retrieved track $r_j^{(q)}$ to be *relevant* if it shares at least one genre with the query track q, and include four definitions of NDCG based on different values of the gain. In the simplest binary case of NDCG_B, we assign $r_j^{(q)}$ a gain of one if it shares at least one genre with q, and zero otherwise. For NDCG_S we assign a gain given by the Szymkiewicz-Simpson coefficient $|G_q \cap G_{r_j^{(q)}}| / \min(|G_q|, |G_{r_j^{(q)}}|)$. For NDCG_J we assign a gain given by the Jaccard coefficient $|G_q \cap G_{r_j^{(q)}}| / |G_q \cup G_{r_j^{(q)}}|$. For NDCG_D we assign a gain given by Sørensen–Dice coefficient $2|G_q \cap G_{r_j^{(q)}}| / (|G_q| + |G_{r_j^{(q)}}|)$. By extending the binary gain, we enforce that a track with a large genre overlap with q leads to a higher NDCG if it is ranked at the top of the list, compared to another with a smaller genre overlap. $\text{NDCG}_{B,S,D,J}$ are aggregated with mean over all retrieval lists.

[4] Although we consider 12 different feature vectors, as listed in Table 1, we report the results of the two unimodal retrieval systems that reached the highest accuracy within each modality, and of multimodal retrieval systems obtained by their combination. We refer the reader to the dashboard for the full results.

As for the studied *beyond-accuracy* metrics, we define the popularity p_m of track m as the sum of its interactions over all users [4] and the *popularity bias B*, i.e., the tendency to retrieve tracks that are more popular than the query track, adapting the method from Lesota et al. [4]: $B = \underset{q \in M}{\text{Median}} \left(\frac{\overline{p_{r(q)}} - p_q}{p_q} \right)$, where $\overline{p_{r(q)}}$ denotes the average popularity of all tracks retrieved for q. A positive B indicates that retrieved tracks are overall more popular than queries.[5] We define *coverage C* as the percentage of all tracks in M that occur in at least one result list for any query [23]. We define *hubness H* as the tendency to often retrieve the same tracks for different queries, leading to non-symmetric results [26, 5, 27, 6]. We measure H in terms of the skewness of the distribution of k-occurrences [5, 6]. We also analyze the *robustness* of unimodal systems, i.e., the extent to which systems based on the same modality (e.g., lyrics) but different representations (e.g., TF-IDF vs. BERT) create similar rankings for the same query, and the influence of each modality in case of multimodal systems, i.e., the *coherence* between results retrieved with unimodal and multimodal systems. We quantify both in terms of Kendall's rank correlation between the lists created by the two retrieval systems to compare, i.e., $\phi^{(\mathbf{f}_{\mathbf{k}_1}, \text{dis}_1)}$ and $\phi^{(\mathbf{f}_{\mathbf{k}_2}, \text{dis}_2)}$.

In the evaluation, NDCG, C, H, and B are computed over all queries $q \in M$ and for $N = 10$ top retrieved tracks.[6] The rank correlations are computed over all queries $q \in M$ and for lists of $|M| - 1$ retrieved tracks, i.e., ranking all tracks apart from the query, since restricting to a shorter list often results in disjoint lists of retrieved tracks.

4. Experiments

In this section, we provide the details on our experimental setup. More specifically, Section 4.1 describes the dataset and the features $\mathbf{f}_{\mathbf{k}}^{(\mathbf{m})}$ representing the content of the music tracks, while Section 4.2 provides details on the setup used to extract the CF representations of the tracks.

4.1. Dataset

Our experiments are based on the Music4All-Onion dataset [28] and its extension released by Peintner et al. [29]. Music4All-Onion is a large-scale multimodal dataset for MIR. We select the tracks for which all the content features are available and that have at least one genre. This results in $|M| = 68{,}641$ tracks. We perform our experiments with nine features for the audio, three for the lyrics, and three for the video modalities, as described in Table 1. For the audio signal, in order to capture short- and long-time dependencies, we consider both the Mel Frequency Cepstral Coefficients (MFCCs), aggregated either with statistical descriptors or as bag-of-audio-words (BoW) computed with `openXBOW` [30], and all the block-level features (BLFs) [11]. We also include the features extracted with Essentia [31] in our analysis, since these include information such as the zero-crossing rate and the attack time, that is complementary to the MFCCs and BLFs. For the lyrics, we consider both statistical representations of word occurrences in terms of TF-IDF, representations obtained with pre-trained instances of word2vec [32], and representations obtained with the `all-mpnet-v2`[7] pre-trained instance of the SentenceTransformer model [33] provided by Hugging Face [34]. We refer to this latter representation of the lyrics as BERT. For the video modality, we consider the visual representations of the YouTube videoclips of the music tracks. These visual representations are obtained by first sampling videoclip frames at 1 fps. The frames are then encoded with pre-trained instances of VGG19 [35], Inception v3 [36], and ResNet [37] and their vector representations are aggregated using max and mean pooling over all frames and for each dimension of the encoding vector, for each track. Finally, the max and mean vectors are concatenated, resulting for each architecture (VGG19, Inception v3, ResNet) in a video feature vector $\mathbf{f}_{\mathbf{k}}$ of double dimensionality with respect to the dimensionality of the visual representation of the pre-trained encoding architecture. We report the results of the two retrieval systems that reached the highest NDCG_J within each modality

[5] Common measures of popularity bias in RSs use the median instead of the mean since it is more robust to outliers.
[6] We refer the reader to the dashboard for the evaluation of retrieval systems on lists of $N = 100$ top retrieved tracks.
[7] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

and refer the reader to the dashboard for the full results. Table 1 summarizes the features $\mathbf{f_k}$ used for each modality, and their corresponding dimensionality n_k.

Modality	Feature $\mathbf{f_k}$	n_k
Lyrics	TF-IDF	994
	Word2vec [32]	300
	BERT [33]	768
Audio	MFCC BoW [30]	500
	MFCC Statistics	104
	Essentia [31]	1,023
	BLF Delta Spectral [11]	1,372
	BLF Correlation [11]	1,326
	BLF Logarithmic Fluctuation [11]	3,626
	BLF Spectral [11]	980
	BLF Spectral Contrast [11]	800
	BLF Variance Delta Spectral [11]	1,344
Videoclip	Inception [36]	4,096
	VGG-19 [35]	8,192
	ResNet [37]	4,096

Table 1
Modality and features $\mathbf{f_k}$ included in our experiments; n_k represents the dimensionality. In the remainder of the paper we report the results of the two features that reached the highest NDCG_J within each modality and refer the reader to the dashboard for the others.

For *multimodal systems* we select the two features that reached the best performance in terms of NDCG_J within each modality (resulting in six features), and consider all possible combinations (resulting in eight combinations for each fusion technique). For *early fusion*, we first normalize the feature vectors to 1 with L_2-norm, and then concatenate them. For *late fusion*, we apply Z-score normalization to the distribution of scores of the individual retrieval systems and then average the normalized scores with weights proportional to NDCG_B.

4.2. Collaborative Filtering Representations

To obtain a representation of the user–item interaction data of each track with each recommendation algorithm (`ItemkNN`, `SVD`, `MultVAE`), we use the set of user–item interactions available in the Music4All dataset [38].[8][9] The characteristics of this set of user-item interactions are summarized in Table 2.

n_{inter}	n_{u}	$n_{\text{t}}^{\text{inter}}$	$n_{\text{t}}^{\text{w/o inter}}$
4,106,678	14,127	67,055	1,586

Table 2
Characteristics of the set of user-item interactions used to obtain the CF item representations with `ItemkNN`, `SVD`, and `MultVAE`. n_{inter} represents the number of user–item interactions, n_{u} the number of users, $n_{\text{t}}^{\text{inter}}$ the number of tracks with at least one interaction, and $n_{\text{t}}^{\text{w/o inter}}$ the number of tracks without interactions.

We set the number of factors in `SVD` to $f = 200$ and the dimension of the latent representation in `MultVAE` to $f = 500$. We fix the batch size to 512, the maximum number of epochs to 300 and apply early stopping with a patience of 10. We set the initial learning rate to 0.003 and reduce the learning

[8] 67,055 out of the 68,641 ($\sim$ 98%) relevant tracks from Music4All have been listened to at least once, i.e., correspond to at least one user–item interaction. Query tracks without user–item interactions lead to a vector of zeros (for `ItemkNN`) or a randomly initialized one (for `SVD` and `MultVAE`), yielding results that are comparable to those of a random retrieval system.
[9] We refer the reader to the dashboard for the results obtained with CF representations based on the set of user–item interactions from the Music4All-Onion dataset, for which 35,702 out of the $|M| = 68{,}641$ ($\sim$ 52%) tracks have been listened to at least once, i.e., correspond to at least one user–item interaction.

rate by a factor of 0.5 if an epoch shows no reduction in the training loss.[10] Following common practice for MRSs [4, 39, 40], we convert the user–item interactions to implicit feedback, binarizing them by setting the entry in the interaction matrix to 1 (positive feedback) if the user listened to the track at least two times, and to 0 otherwise.

5. Results

In this section, we analyze the results of our experiments on music retrieval systems. Section 5.1 compares the performance of uni- and multi-modal retrieval systems in terms of accuracy, coverage C, hubness H and popularity bias B defined as described in Section 3.2. Section 5.2 analyzes the robustness of each content modality (audio, lyrics, videoclips) and the impact of each modality on multi-modal systems. In the evaluation, NDCG, C, H, and B are computed over 68,641 queries and for top 10 retrieved tracks. The rank correlations are computed over 68,641 queries and for lists of 68,640 retrieved tracks, i.e., ranking all tracks, since restricting to a shorter list often results in disjoint lists of retrieved tracks. Since all retrieval systems reached higher $\text{NDCG}_{B,S,D,J}$ with cosine similarity, we restrict our discussion to retrieval systems based on cosine.

5.1. Accuracy, Coverage, Hubness, and Popularity Bias

Table 3 shows the $\text{NDCG}_{B,S,D,J}$, hubness H, coverage C and popularity bias B of the retrieval systems. As baseline for comparison, the first block of the table shows the results of a system retrieving tracks at random for each query. The following block refers to retrieval systems based on one feature, either from one content modality (lyrics, audio, or video)[11] or from CF representations. The last block refers to multimodal retrieval systems based on all content modalities, either with early- or with late-fusion. As described and motivated in Section 3.2, $\text{NDCG}_{B,S,D,J}$ show the mean and B the median over all queries. For these metrics, all differences between the best performing system in each sub-block (in bold) and the remaining ones are statistically significant ($p < 0.05$ for paired t-tests using Bonferroni correction to account for multiple comparisons), aside from those between BLF and ResNet. We first observe that within content-based retrieval systems, video features lead to the highest accuracy, especially when measured with $\text{NDCG}_{S,D,J}$. The fact that audio features are competitive in terms of NDCG_{B} but reach a worse performance in terms of $\text{NDCG}_{S,D,J}$ indicates that both audio and videoclips are comparable in retrieving tracks that share *at least one genre*, but videoclips lead to results that share *more genres* with the query tracks. Among content-based retrieval systems, fusion techniques generally tend to reach higher accuracies than systems based on individual modalities, with early-fusion leading to higher $\text{NDCG}_{B,S,D,J}$ compared to late-fusion. `ItemkNN` reaches the highest $\text{NDCG}_{B,S,D,J}$ and all content-based retrieval systems are outperformed by all CF systems. This shows that collaborative data, which do not include any explicit information on the track content, are also useful for MIR tasks beyond recommendation. This higher accuracy, however, comes at the cost of a higher hubness and an overall tendency to a higher popularity bias (aside from `MultVAE`). Surprisingly, however, CF systems also outperform content-based ones in terms of coverage. This indicates their tendency to retrieve different, but more popular, tracks. We hence conclude that if accuracy and coverage are to be prioritized when retrieving music, it is in the interest of the MIR system provider to select CF representations. However, these are not always available, e.g., on platforms where interaction data are not collected. In that case, multimodal systems should be preferred.

[10] We use default hyperparameters since any data split leading to a reasonable optimization of the MRSs would not be meaningful for the retrieval system. For instance leaving out a set of tracks for validation would lead to an embedding dimensionality that is not optimal when all tracks are considered, while a split at the interaction or user level would be prone to information leakage, since the same tracks would be selected for the hyperparameter optimization and evaluation.

[11] For unimodal content-based retrieval systems, we report the results of the two features that reached the highest NDCG_{J} within each modality and refer the reader to the dashboard for the others.

5.2. Robustness and Feature Impact

Figure 1 shows the rank correlations between pairs of content-based retrieval systems. The systems are divided into unimodal, early-, and late-fusion systems by orange dashed lines. E and L represent early- and late-fusion and the first index refers to the audio, the second to the lyrics, and the last to the video feature, respectively. The orange bold numbers represent the average over the corresponding sub-blocks, excluding the ones on the main diagonal. We first observe that all correlations are positive, which indicates that retrieval systems do not invert the order of results, not even across modalities. For unimodal retrieval systems (block $U \times U$), the correlations are typically close to zero across modalities and close to 0.5 between features of the same modality. This is especially true for audio and video, indicating that they are more robust under a change of the representation, in contrast to lyrics. The average correlation between multimodal systems is comparable within and across early- and late-fusion systems (0.63 for $E \times E$ and $L \times L$; 0.65 for $E \times L$), and reaches values close to 1 when more than one same feature is leveraged by both (e.g., between L_MBI and E_MBI). This indicates that multimodal systems are affected by the representation of the modalities more than by the aggregation technique. In fact, the choice of a particular fusion technique only marginally affects the retrieval results. This observation is of practical relevance for system providers, especially in cases where the system infrastructure might rule out certain fusion techniques. The correlations between unimodal and multimodal systems (entries in blocks $E \times U$ and $L \times U$) are higher if the feature is shared than if it is not; Within those cases, for early-fusion the correlations are higher with unimodal systems based on video than on other modalities, while for late-fusion those are close to each other. This indicates that multimodal systems based on late-fusion leverage information from each modality in a more balanced way compared to early-fusion systems. This observation, together with the one concerning the high correlation of early- and late-fusion systems sharing the same features, and their comparable performance in terms of accuracy is of particular interest for MIR system providers in cases in which they want to reflect all modalities, instead of prioritizing one over the others.

6. Conclusions

This work compared the accuracy, coverage, hubness, popularity bias, and robustness of similarity-based music retrieval systems based on content or collaborative data, as well as the coherence between unimodal and multimodal systems. The results provide useful information to platform providers, especially in cases where the choice of modality or fusion technique has to consider aspects beyond accuracy, or in which one or more representations of the music tracks are missing. One noteworthy finding is the very good accuracy of ResNet features from videoclips, considering they are computed from the image content only, and disregarding the actual music audio content. This surprising result might be originating from the genre-based evaluation setting, and could indicate that music tracks of a same genre share distinctive visual characteristics (e.g., , videoclips for emo rock songs are often filmed in black and white). Our definition of relevance is framed as finding tracks of the same music genres of a query track; this constitutes one limitation of the current work. Future work could extend the evaluation to other evaluation settings, e.g., framing the evaluation as playlist completion given a seed track. These evaluations, taken together with the current one, would provide a more comprehensive view on the impact of content features on MIR tasks. Another limitation of our work is that although we included representations of lyrics, videoclips, and collaborative data based on a NN, we only used hand-crafted features for the audio signal. The reason is that many (deep) NNs for music are pre-trained on tags or genres. The learned models would therefore be prone to information leakage, considering our relevance definition. Additionally, it would be interesting to compare the accuracy and beyond-accuracy metrics reported in this work with those actually perceived by users, e.g., via user studies. We leave these analyses for future work.

		$NDCG_B \uparrow$	$NDCG_S \uparrow$	$NDCG_D \uparrow$	$NDCG_J \uparrow$	$H \downarrow$	$C \uparrow$	$B \downarrow$
	Random	0.4459	0.1762	0.1198	0.0833	0.3213	0.9999	1.8250
Lyrics	*TF-IDF*	0.5229	0.2282	0.1570	0.1126	5.2423	0.7542	1.8159
	BERT	0.5802	0.2760	0.1942	0.1421	13.1569	0.8235	1.8980
Audio	*MFCC*	0.6096	0.3014	0.2172	0.1619	3.5705	**0.8958**	1.8444
	BLF	**0.6136**	0.3072	0.2221	0.1661	**3.5074**	0.8486	**1.7267**
Videoclip	*Inception*	0.6052	0.3211	0.2567	0.2055	10.9384	0.8259	1.8538
	ResNet	0.6119	**0.3294**	**0.2636**	**0.2116**	6.0527	0.8857	1.8406
CF	`ItemkNN`	**0.7422**	**0.4936**	**0.4172**	**0.3516**	65.2356	**0.9481**	1.9167
	`SVD`	0.7233	0.4400	0.3639	0.2978	**55.6481**	0.9202	1.8889
	`MultVAE`	0.7161	0.4502	0.3709	0.3040	74.9379	0.9011	**1.7875**
Early f.	*BLF, BERT, Inception*	0.6656	0.3567	0.2717	0.2119	11.6694	0.7453	1.8250
	BLF, BERT, ResNet	0.6807	0.3784	0.2957	0.2350	10.3243	0.8125	1.8426
	BLF, TF-IDF, Inception	0.5941	0.2832	0.2050	0.1534	6.5312	0.7762	1.8286
	BLF, TF-IDF, ResNet	0.6234	0.3206	0.2440	0.1904	6.1970	**0.8703**	1.8108
	MFCC, BERT, Inception	0.6687	0.3602	0.2735	0.2133	10.1415	0.7421	**1.8000**
	MFCC, BERT, ResNet	**0.6839**	**0.3820**	**0.2975**	**0.2364**	10.1745	0.8076	1.8167
	MFCC, TF-IDF, Inception	0.5968	0.2862	0.2068	0.1549	6.2216	0.7761	1.8500
	MFCC, TF-IDF, ResNet	0.6259	0.3229	0.2453	0.1914	**5.7451**	0.8650	1.8167
Late f.	*BLF, BERT, Inception*	0.6694	0.3689	0.2843	0.2236	11.1819	0.7426	**1.7907**
	BLF, BERT, ResNet	0.6794	0.3741	0.2894	0.2285	9.4432	**0.7999**	1.8348
	BLF, TF-IDF, Inception	0.5723	0.2646	0.1870	0.1370	6.2305	0.7413	1.8231
	BLF, TF-IDF, ResNet	0.5790	0.2685	0.1903	0.1399	5.9580	0.7899	1.8105
	MFCC, BERT, Inception	0.6710	0.3711	0.2852	0.2244	9.9828	0.7429	1.7975
	MFCC, BERT, ResNet	**0.6812**	**0.3768**	**0.2908**	**0.2296**	9.9276	0.7979	1.8000
	MFCC, TF-IDF, Inception	0.5708	0.2640	0.1858	0.1360	6.1250	0.7319	1.8455
	MFCC, TF-IDF, ResNet	0.5774	0.2677	0.1891	0.1389	**5.7306**	0.7780	1.8333

Table 3
Accuracy, hubness H, coverage C, and popularity bias B of the systems. For accuracy, we report $NDCG_{B,S,D,J}$, i.e., using binary, Szymkiewicz-Simpson, Sørensen–Dice, or Jaccard as gain. Random refers to a baseline system retrieving tracks at random for each query. The following block refers to retrieval systems based on one feature, either content or CF. The last block refers to early- or late-fusion retrieval systems. All differences between the best performing system for unimodal, CF, early- and late-fusion, i.e., the best performing system in each sub-block (in bold), and the others are statistically significant ($p < 0.05$ for paired t-tests using Bonferroni correction to account for multiple comparisons), apart from the differences between BLF and ResNet.

7. Declaration on Generative AI

No generative AI tool was used during the preparation of this work.

8. Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF) https://doi.org/10.55776/P33526, https://doi.org/10.55776/DFH23, https://doi.org/10.55776/COE12, https://doi.org/10.55776/P36413.

References

[1] M. Schedl, A. Flexer, J. Urbano, The neglected user in music information retrieval research, J. Intell. Inf. Syst. 41 (2013) 523–539.
[2] M. Schedl, H. Zamani, C. Chen, Y. Deldjoo, M. Elahi, Current challenges and visions in music recommender systems research, International Journal of Multimedia Information Retrieval 7 (2018) 95–116.
[3] M. Schedl, P. Knees, B. McFee, D. Bogdanov, Music recommendation systems: Techniques, use cases, and challenges, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, 2022, pp. 927–971.

Figure 1: Kendall's rank correlations between pairs of content-based retrieval systems. E and L represent early- and late-fusion; the first index refers to audio, the second to lyrics, and the last to video features, respectively. For instance, E_BTI represents the early-fusion based on the BLF spectral representation of the audio, TF-IDF of the lyrics, and Inception of the videoclip. Orange dashed lines separate unimodal, early-, and late-fusion. Orange bold numbers represent the average correlation over the corresponding sub-blocks. The diagonal values of the shown lower triangular matrix are excluded from the computations.

[4] O. Lesota, A. Melchiorre, N. Rekabsaz, S. Brandl, D. Kowald, E. Lex, M. Schedl, Analyzing item popularity bias of music recommender systems: Are different genders equally affected?, in: Proc. of ACM RecSys, 2021, pp. 601–606.
[5] D. Schnitzer, A. Flexer, M. Schedl, G. Widmer, Local and global scaling reduce hubs in space, Journal of Machine Learning Research 13 (2012) 2871–2902.
[6] K. Seyerlehner, A. Flexer, G. Widmer, On the limitations of browsing top-n recommender systems, in: Proc. of ACM RecSys, 2009, p. 321–324.
[7] T. Akama, H. Kitano, K. Takematsu, Y. Miyajima, N. Polouliakh, Auxiliary self-supervision to metric learning for music similarity-based retrieval and auto-tagging, PLOS ONE 18 (2023) 1–20.
[8] A. C. M. da Silva, D. F. Silva, R. M. Marcacini, Multimodal representation learning over heterogeneous networks for tag-based music retrieval, Expert Systems with Applications 207 (2022) 117969.
[9] Y. Deldjoo, M. Schedl, P. Knees, Content-driven music recommendation: Evolution, state of the art, and challenges, Computer Science Review 51 (2024) 100618.
[10] H. Eghbal-Zadeh, B. Lehner, M. Schedl, G. Widmer, I-vectors for timbre-based music similarity and music artist classification, in: Proc. of ISMIR, 2015, pp. 554–560.
[11] K. Seyerlehner, G. Widmer, M. Schedl, P. Knees, Automatic music tag classification based on block-level features, in: Proc. of SMC, 2010.
[12] P. Knees, M. Schedl, A survey of music similarity and recommendation from music context data, ACM Trans. Multimedia Comput. Commun. Appl. 10 (2013).
[13] M. Won, S. Oramas, O. Nieto, F. Gouyon, X. S. Serra, Multimodal metric learning for tag-based music retrieval, in: Proc. of IEEE ICASSP, 2021, p. 591–595.
[14] J. Lee, N. Bryan, J. Salamon, Z. Jin, J. Nam, Metric learning vs classification for disentangled music representation learning, in: Proc. of ISMIR, 2020, pp. 439–445.
[15] C. Thomé, S. Piwell, O. Utterbäck, Musical audio similarity with self-supervised convolutional neural networks, in: Proc. of ISMIR, 2022, p. LBR & Demo Papers.
[16] P. Manocha, Z. Jin, R. Zhang, A. Finkelstein, Cdpam: Contrastive learning for perceptual audio similarity, in: Proc. of IEEE ICASSP, 2021, pp. 196–200.
[17] A. Ferraro, J. Kim, S. Oramas, A. Ehmann, F. Gouyon, Contrastive learning for cross-modal artist retrieval, in: Proc. of ISMIR, 2023.
[18] L. Su, C.-C. M. Yeh, J.-Y. Liu, J.-C. Wang, Y.-H. Yang, A systematic evaluation of the bag-of-frames representation for music information retrieval, IEEE Transactions on Multimedia 16 (2014) 1188–1200.
[19] C. Plachouras, P. Alonso-Jiménez, D. Bogdanov, mir_ref: A representation evaluation framework for music information retrieval tasks, in: Proc. of Machine Learning for Audio Workshop co-located with NeurIPS, New Orleans, LA, USA, 2023.
[20] M. Ge, C. Delgado-Battenfeld, D. Jannach, Beyond accuracy: Evaluating recommender systems by coverage and serendipity, in: Proc. of ACM RecSys, 2010, pp. 257–260.
[21] M. Kaminskas, D. Bridge, Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems, ACM Transactions on Interactive Intelligent Systems 7 (2016).
[22] V. W. Anelli, A. Bellogín, T. Di Noia, C. Pomo, Reenvisioning the comparison between neural collaborative filtering and matrix factorization, in: Proc. of ACM RecSys, 2021, pp. 521–529.
[23] A. Gunawardana, G. Shani, S. Yogev, Evaluating recommender systems, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, 2022, pp. 547–602.
[24] M. Radovanović;, A. Nanopoulos, M. Ivanović;, Hubs in space: Popular nearest neighbors in high-dimensional data, Journal of Machine Learning Research 11 (2010) 2487–2531.
[25] D. Liang, R. G. Krishnan, M. D. Hoffman, T. Jebara, Variational autoencoders for collaborative filtering, in: Proc. of ACM WWW, 2018, pp. 689–698.
[26] A. Flexer, D. Schnitzer, J. Schlüter, A mirex meta-analysis of hubness in audio music similarity, in: Proc. of ISMIR, 2012, pp. 175–180.
[27] A. Flexer, M. Dörfler, J. Schlüter, T. Grill, Hubness as a case of technical algorithmic bias in music

recommendation, in: Proc. of ICDMW, 2018, pp. 1062–1069.
[28] M. Moscati, E. Parada-Cabaleiro, Y. Deldjoo, E. Zangerle, M. Schedl, Music4all-onion - A large-scale multi-faceted content-centric music recommendation dataset, in: Proc. of ACM CIKM, 2022, pp. 4339–4343.
[29] A. Peintner, M. Moscati, Y. Kinoshita, R. Vogl, P. Knees, M. Schedl, H. Strauss, M. Zentner, E. Zangerle, Nuanced music emotion recognition via a semi-supervised multi-relational graph neural network, Transactions of the International Society for Music Information Retrieval (2025).
[30] M. Schmitt, B. Schuller, Openxbow: Introducing the passau open-source crossmodal bag-of-words toolkit, Journal of Machine Learning Research 18 (2017) 3370–3374.
[31] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. R. Zapata, X. Serra, Essentia: An audio analysis library for music information retrieval, in: Proc. of ISMIR, 2013, pp. 493–498.
[32] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proc. of ICLR, 2013.
[33] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proc. of EMNLP, 2019, pp. 3973–3983.
[34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proc. of EMNLP System Demos, 2020, pp. 38–45.
[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. of ICLR, 2015.
[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. of IEEE CVPR, 2015, pp. 1–9.
[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. of IEEE CVPR, 2016, pp. 770–778.
[38] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim, M. A. Domingues, et al., Music4all: A new music database and its applications, in: Proc. of IEEE IWSSIP, 2020, pp. 399–404.
[39] A. B. Melchiorre, N. Rekabsaz, C. Ganhör, M. Schedl, Protomf: Prototype-based matrix factorization for effective and explainable recommendations, in: Proc. of RecSys, 2022, p. 246–256.
[40] A. B. Melchiorre, N. Rekabsaz, E. Parada-Cabaleiro, S. Brandl, O. Lesota, M. Schedl, Investigating gender fairness of recommendation algorithms in the music domain, Information Processing & Management 58 (2021) 102666.