Unsupervised Cross-Modal Audio Representation Learning from Unstructured Multilingual Text

Alexander Schindler
Center for Digital Safety and Security
Austrian Institute of Technology
Vienna, Austria
alexander.schindler@ait.ac.at

Sergiu Gordea
Center for Digital Safety and Security
Austrian Institute of Technology
Vienna, Austria
sergiu.gordea@ait.ac.at

Peter Knees
Faculty of Informatics
TU Wien
Vienna, Austria
peter.knees@tuwien.ac.at

ABSTRACT

We present an approach to unsupervised audio representation learning. Based on a triplet neural network architecture, we harnesses semantically related cross-modal information to estimate audio track-relatedness. By applying Latent Semantic Indexing (LSI) we embed corresponding textual information into a latent vector space from which we derive track relatedness for online triplet selection. This LSI topic modelling facilitates fine-grained selection of similar and dissimilar audio-track pairs to learn the audio representation using a Convolution Recurrent Neural Network (CRNN). By this we directly project the semantic context of the unstructured text modality onto the learned representation space of the audio modality without deriving structured ground-truth annotations from it. We evaluate our approach on the Europeana Sounds collection and show how to improve search in digital audio libraries by harnessing the multilingual meta-data provided by numerous European digital libraries. We show that our approach is invariant to the variety of annotation styles as well as to the different languages of this collection. The learned representations perform comparable to the baseline of handcrafted features, respectively exceeding this baseline in similarity retrieval precision at higher cut-offs with only 15% of the baseline's feature vector length.

CCS CONCEPTS

• **Information systems** → **Content analysis and feature selection**; **Multilingual and cross-lingual retrieval**; **Speech / audio search**;

KEYWORDS

audio representation learning, cross-modal learning, deep neural networks

1 INTRODUCTION

Audio representations aim to capture intrinsic properties and characteristics of the audio content to facilitate complex tasks such as classification (acoustic scenes [6, 16], music genres [15]), regression (emotion recognition [31]) or similarity estimation (music,[13] general audio [22]). In the context of this paper we focus on their application in audio similarity estimation and retrieval. More specifically, for heterogeneous audio collections provided by digital libraries in the cultural heritage domain.

Digital libraries (DL) present unique challenges for information retrieval research. The information need of DL users is highly specific and users are often highly experienced within the search domain. The challenging requirement for effective tools to search and discover content in large databases faces major obstacles such as heterogeneity, multi-modality and often multi-linguality of the content stored in these databases. The common approach taken to satisfy these information needs is to provide as much, rich and accurate meta-data as possible and to apply information retrieval and semantic computing [20] technologies to index this meta-data to facilitate efficient search. Either of these approaches requires considerable amounts of manual interaction to annotate or interlink data items and does not scale well. Approaches based on meta-data require that DL users are acquainted with the correct terminology used to describe the item or the categories. Based on the fact that many collections in DLs are aggregated and curated from scientists from specific research disciplines such as history, archaeology or musicology, this terminology can be very specific and not everyone might be familiar with it on the same level. Further, not every type of information can be efficiently described using textual meta-data. Content related relations such as "sounds like" are highly complex and difficult to describe by meta-data. This heterogeneity is also a challenge for the definition and modeling of the *acoustic similarity* function. The Europeana[1] Sounds data-set [23] contains besides *Music* also *Spoken Word* in form of interviews, radio news broadcast, public speeches, field recorded *Animal-* and *Ambient-* or *Environmental-Sounds*. Additionally, the recordings vary in instrumentation and recording quality (from digitized wax-tapes to born-digital content). In [23] we applied a diverse set of handcrafted audio- and music-descriptors to model an audio-content based similarity estimation function for the *Europeana* data. The most critical part of this approach was the selection of the features to adequately describe the heterogeneous semantics of the different collections in the data-set, as well as the balancing of the feature weights to approximate the subjective similarity estimation. Feature weight
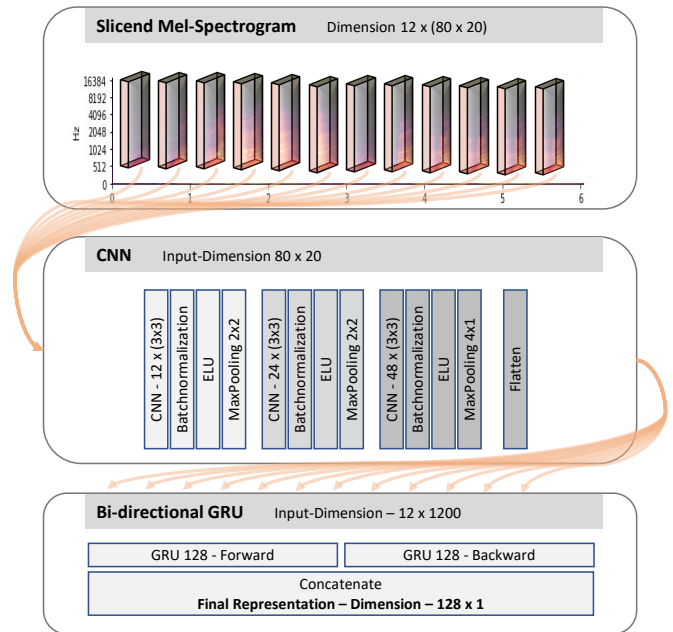
---

[1]https://www.europeana.eu

optimization was approached empirically through a predefined set of similar records. During an iterative process the weights of the different features were adapted. This manual optimization process is sub-optimal in terms that it only optimizes towards a small set of manually selected items. An optimization against the entire dataset would require that pairwise similarity estimations would be available as ground-truth-assignments. Because creating such assignments is not feasible on large scales, a common approach is to define similarity by categorical membership or identity. In [19] the authors defined music similarity on tracks originating from the same artists and used a triplet deep neural network architecture (see Section 1) to learn an optimized music optimization. Using identity or categorical data as ground-truth is sub-optimal [24] because it usually defines acoustic similarity too coarse. In [24] we learned a music representation from multi-label assignments using Latent Semantic indexing (LSI) to project the discrete information into a continuous space from which a track-similarity function based on tag-relatedness was derived. This tag-based track-similarity was then transferred to the audio space using a Triplet Network with a margin maximizing loss and online triplet selection strategy. Following this approach a network learns to maximize the margin between the distances of a reference track and its positive similar example and its negative dissimilar example, in a learned semantic embedding space, based on the constraint, that the vector distance of the positive pair should be much smaller than the distance of the negative pair. The challenge is the selection of positive and negative examples to a reference track. Still, these approaches rely on ground-truth assignments for supervised learning of audio representations.

In this paper we build on the conclusions of [24] that by adding more content related information to the input space of the LSI projections, the definition of LSI topics improves, resulting in a better track-relatedness function. We adopt this approach and extend it from a discrete categorical space with a fixed vocabulary to an open, unstructured, multi-lingual free-text space. The metadata provided by professional librarians contains such item-related descriptions including descriptions of audible content. The major contribution of this approach is, that it uses unstructured text to derive track-relatedness and does not require structured ground-truth assignments. To demonstrate our approach we first discuss and position our approach within related work in Section 2 before we describe our method in Section 3. We extensively evaluate the approach in Section 4 and thoughtfully discuss the results in Section 5 before we draw conclusion and discuss future work in Section 6.

2 RELATED WORK

2.1 Content Retrieval in Music Digital Libraries

Concerning Music Digital Libraries (DL) this paper is mostly related to [23] where we presented an approach to audio-content similarity estimation within Europeana Sounds project. Following a late-fusion approach, the weighted combination of different audio-content descriptors was applied to highly heterogeneous data. The presented evaluation method is also adopted in this paper. Issues of data aggregation in audio DLs are described in [30].

Figure 1: Model Overview: Convolutional Recurrent Neural Network (CRNN). a) the input Mel-Spectrogram (80x130) is split into 12 segments (80x12, 50% overlap). b) each segment is processed by a shared CNN stack resulting in a sequence of 12 intermediate embedding vectors. c) the sequence is processed by a bi-directional GRU and its output is the learned representation (256x1).

2.2 Music Similarity Retrieval

Search-by-example, such as finding music tracks that are similar to a query track, is an actively researched task [11, 13]. Research on music similarity estimation currently faces two major obstacles. First, music similarity is a highly subjective concept and is strongly influenced by the listening habits and music taste of the listener [1]. Second, state-of-the-art approaches to music similarity estimation are still not able to satisfactorily close the semantic gap between the computational description of music content and the perceived music similarity [10]. The many facets of music similarity - such as concrete music characteristics (e.g. rhythm, tempo, key, melody, instrumentation), perceived mood (e.g. calm, aggressive, happy), listening situation (e.g. for dinner, to concentrate, for work out), musicological factors (e.g. composer influenced by) - complicate the definition of a unified music description which captures all semantic music concepts. Traditionally this has been approached by defining a set of features, which extract certain low level music characteristics such as timbre [17] or rhythm [14], mid-level properties such as chords [18], but also high-level features. This approach faces the problem that hand-crafted feature design is neither scaleable nor sustainable [9]. Representation learning using Deep Neural Networks (DNN) has been actively explored in recent years [27, 28] as an alternative to feature engineering. Although some of these approaches outperform feature-based methods, a major

obstacle is their dependency on large amounts of training data. Although it has been shown that shallow DNNs have an advantage on small datasets [25] they struggle to describe the latent complexity of music concepts and do not generalize on large datasets [9].

2.3 Representation Learning (RL)

RL using DNNs gained attention through the publication of *FaceNet* [27] which significantly improved the state-of-the-art of face re-identification. This approach is based on global item relatedness where faces are similar when they belong to the same person and otherwise are dissimilar. A similar approach using the global relatedness of performing artists has been applied to music data [19]. Contextualized relatedness especially in the domain of music has been used in [21]. A similar approach to this paper of estimating tag-relatedness from user-tags was taken in [21]. Latent Dirichlet Analysis (LDA) was used to project the categorical data into a numerical space. The approach was evaluated using a siamese neural network on three smaller datasets including the MSD subset using the noisy user-generated tag-sets of the Last.fm dataset. A differentiated evaluation of the learned semantic context as provided in this paper was missing. Concerning Representation Learning the presented paper in mostly related to [24] in which we extended the Million Song Dataset (MSD) [2] with additional ground truth multi-label assignments for *Moods*, *Styles* and *Themes*. Further, we extended the single-label *Genre* labels provided in [26] to multi-label assignments. In [24] a music representation was learned from these multi-label assignments using Latent Semantic indexing (LSI) to project the categorical information into a continuous space from which a track-similarity function based on tag-relatedness was derived. This tag-based track-similarity was then transferred to the audio space using a Triplet Network with a margin maximizing loss and online triplet selection strategy.

3 METHOD

The proposed method is based on a triplet neural network architecture to learn the contextualized semantic representation using a max-margin hinge loss with online triplet selection.

3.1 Representation Learning

To learn the acoustic representation we use a triplet network based on a shared Convolutional Recurrent Neural Network (CRNN) architecture [3]. The base-model is described in Section 4.1 and depicted in Figure 1. Using this triplet network, an input audio spectrogram x is embedded $f(x_i^a)$ into a d-dimensional Euclidean space $\mathbb{R}^d$. The input consists of a triplet of audio content items: a query track (anchor) x_i^a, a track similar (positive) x_i^p and one dissimilar (negative) x_i^n to the query. The objective is to satisfy the following constraint:

$$\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 + \alpha < \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 \qquad (1)$$

For $\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \tau$, where $\left\| f(x_i^a) - f(x_i^p) \right\|_2^2$ is the squared Euclidean distance between x_i^a and x_i^p, which should be much smaller than the distance between x_i^a and x_i^n. α is the enforced margin between positive and negative pair-distances. τ represents
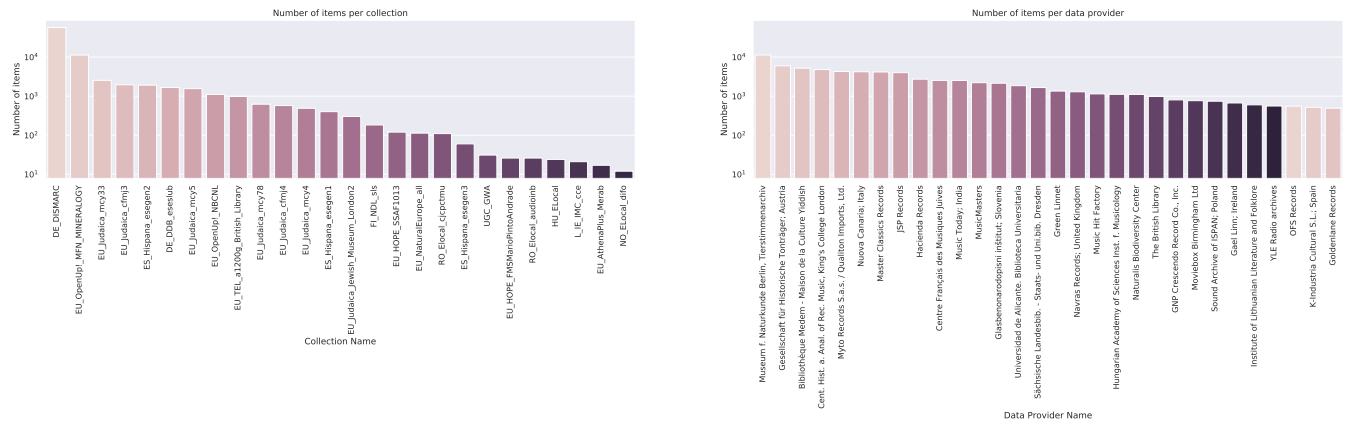
the set of all possible triplets in the training-set. The objective of Eq. 1 is reformulated as the following triplet-loss function:

$$\sum_{i=1}^{N} \max \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \alpha \right] \qquad (2)$$

3.1.1 Online Triplet Selection. Efficient selection of triplets is a crucial step in training the network. Generating all possible triplet combinations τ is inefficient due to the cubic relation and the lacking contribution to the training-success of triplets not violating Eq. 1. Thus it is required to select hard triplets violating this constraint. A common approach to this is *online triplet selection* where triplets are combined on a mini-batch basis [27]. To select appropriate triplets the batch size needs to be appropriately large. We use a batch size of 400 tracks. Their corresponding log-scaled Mel-Spectrograms (see Sec. 4.4) are embedded into a latent space. The selection of positive and negative examples is based on the semantically embedded textual information extracted from the meta-data (explained in more detail in Sec. 3.2). The pairwise cosine-distance $\cos(LSI_1^{ts}, LSI_2^{ts})$ matrix of the corresponding l_2 normalized LSI-embeddings is calculated for all mini-batch instances. The diagonal elements are set to zero to avoid identical pairs. Thresholds for pair-selection were evaluated empirically by analyzing the distribution of the cosine-distance space of the LSI-embeddings and set to $\cos(LSI_1^{ts}, LSI_2^{ts}) \geq 0.8$ (upper) and $\cos(LSI_1^{ts}, LSI_2^{ts}) < 0.5$ (lower). For each row in the LSI-embeddings similarity matrix that contains valid positive and negative instances, the squared Euclidean distances of the corresponding audio embeddings are calculated. *argmin* is computed to identify relevant positive and negative pairs. This deviates from the original approach [27] where *argmin* is used to select hard negatives and *argmax* for hard positive pair examples. The intention of the approach presented in [27] is to be be invariant to image background as well as to changes in pose, color and illumination. This is supported by their hard triplet selection method which enforces to learn highly discriminate object-related features. The "sounds like" audio similarity of in this paper defines a global similarity also taking "background noise" into account. This is emphasized by using *argmin* to select tracks which similar features in the embedding space as positive pairs. Finally, instances where no positive and negative example are found are removed.

3.2 Relatedness Measure

To train the triplet network with target values reflecting the similarity of two records, we build a measure capturing the similarity of their associated meta-data descriptions. To this end, we make use of all the meta-data entries in all categories (see Fig. 3) and apply text processing to their concatenation, i.e. a standard word-level *term frequency by inverse document frequency* (TFIDF) scheme emphasizing terms specific to individual record meta-data in favor of meta-data entries common to many records. This should lower the impact of collection specific keywords on the learned audio representation while not fully discarding this underlying relationship. To combat the facts that meta-data entries are often correlated and/or applied inconsistently across different collections, i.e., effects similar to *synonymy* and *polysemy* in natural language processing, respectively, as a next step, we perform Latent Semantic

Figure 2: Data-set overview: left) Number of items per collection. right) number of items per data provider. Y-axis is log-scaled for both charts due to unbalanced frequency distributions.

Indexing (LSI) [4]. LSI models latent topics as semantic associations of terms in a corpus of documents (in our case the individual words over all metadata of all records weighted according to the TFIDF scheme). Technically, LSI operates on the $m \times n$ weight matrix W with $w_{ij} = tfidf_{ij}$, where each row corresponds to a term $t_{j=1\ldots m}$ and each column to a record $a_{i=1\ldots n}$. Latent topics—and with it an implicit spectral clustering—are derived by performing truncated singular value decomposition (SVD) for approximation of W. As a result, each individual record is represented in the LSI-derived concept space via a cluster affinity vector (in the following referred to as LSI vector), cf. [12]. A characteristic of SVD is that emerging topics are sorted in order of decreasing importance wrt. reconstruction of the data. Therefore, the number of considered topics l (referring to the first l dimensions of the LSI vectors) can be used to steer the trade-off between generalization of the model and preservation of the original meta-data information in W.

For calculation of record relatedness using the meta-data, we calculate the cosine similarity between the records' LSI vectors. From this, we sample positive and negative examples to be presented to the triplet network.

4 EVALUATION

The aim of this evaluation is to asses if our method facilitates to harness semantic information from the meta-data space to learn a corresponding, general applicable audio representation from it. To show this, we perform three experiments using the task of audio similarity retrieval with the following settings:

(1) **Baseline:** selection of weighted handcrafted features [25] intended to show how neural network based approaches compare against state-of-the-art handcrafted feature-sets reported in literature.

(2) **Track-relatedness by collection:** triplet-based neural network using collection membership for online triplet selection. Acts as a second baseline, representing audio representation learning approaches relying on categorical data for triplet-selection. We therefor use collection membership to select positive and negative track-pairs.

(3) **Track-relatedness by LSI similarity:** our approach - triplet-based neural network using LSI-vector similarity for online triplet selection.

We perform controlled experiments. The same model architecture as described in the following subsection is used for all experiments. We further take control over all random processes such as kernel initialization, shuffling of training instances after each epoch to reduce random effects and variance of the experimental results. The same training, validation and test splits are used in all experiments. By controlling all these parameters to our best knowledge we hypothesize that the learned representations are only influenced by LSI representation of the semantic space of the textual meta-data.

4.1 Model Architecture

For the evaluation we are using a Convolutional Recurrent Neural Network (CRNN) [3, 22]. A CRNN is a combination of a Convolutional Neural Network (CNN) stack and a Recurrent Neural Network (RNN). The CNN learns to identify patterns in the local 2D space of the input Spectrograms. The resulting feature transformation is passed on to the RNN which identifies sequential patterns in this intermediate embedding space. In our model the context learned by the RNN represents the final learned audio representation.

The model concept and architecture is depicted in Figure 1. Instead of globally pre-normalizing the input-space, we use a Batch-Normalization layer on top of our model.This normalized input matrix of shape 80x130 is then split into 12 sequential segments of shape 80x20 which overlap by 50% (see top of Fig. 1). Each segment is then processed by the CNN stack which consists of three blocks - each one containing a convolution layer with 3x3 filter units, BatchNormalization followed by an Exponential Linear Unit (ELU) activation function and MaxPooling to down-sample the feature-maps. For the specific parametrization of the layers please refer to Fig. 1). The feature-maps of the last block are flattened to a feature vector representing the intermediate non-linear feature transformation. The concatenation of all 12 feature vectors serves as input to a bi-directional Gated Recurrent Unit (GRU) which learns

Table 1: Overview of baseline audio representation. Audio-content *descriptors* with corresponding acoustic *categories* and feature *weights* as well as the cumulative *category weight* optimized for *Europeana Sounds* evaluation data.

Category	Feature	Description	W_{feat}	W_{cat}
Timbre	**MFCC**	Timbre [29]	23%	
	SSD	Spectral desc. [14]	8%	39%
	SPEC CENT	Pitch [29]	8%	
Rhythm	**RP**	Rhythm [14]	18%	
	BPM	Tempo [5]	7%	25%
Harmony	**CHROMA**	Harmonic Scale [29]	12%	
	TONNETZ	Harmonic desc. [7]	12%	24%
Loudness	**RMSE**	Loudness [29]	9%	9%
Noise	**ZCR**	Noisiness [29]	3%	3%

sequential patterns in the feature space. The resulting context of the GRU is used as model output - which has 256 dimensions - and represents the learned audio representation.
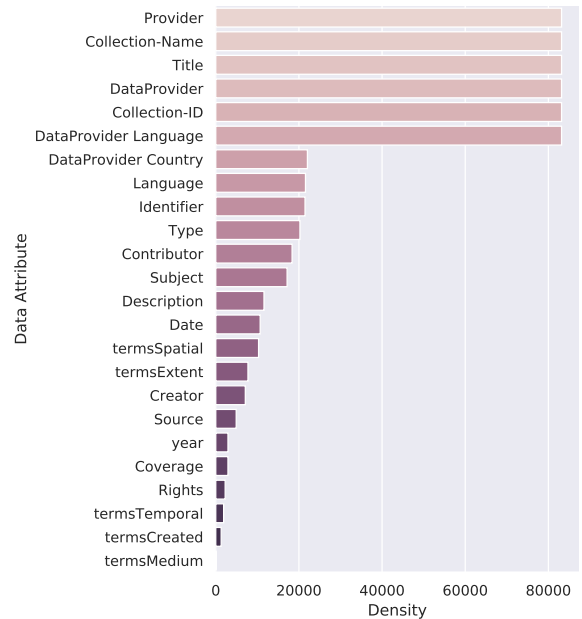
4.2 Baseline Architecture

The baseline approach is based on a selection of handcrafted music- and audio features. This feature-set has been specifically designed and optimized for the *Europeana Sounds* audio collection [25]. Audio simlarity is calculated by late fusing the different features using different similarity metrics. The final similarity is defined the the sum of the weighted distance space. Table 1 depicts the feature-set defined in [25]. Features are selected to describe five acoustic and musicological concepts of the heterogeneous semantics of the different collections in the *Europeana Sounds* data-set. The balancing of the feature weights aims to approximate the subjective similarity.

For the baseline experiments we extract and aggregate the features listed in Table 1 according the procedure described in [25]. To harmonize the evaluation of this paper we transformed the model to an early-fusion approach. Thus, we first standardised the value-spaces of all feature-sets separately. Then, we normalize each feature-set according their dimensionality to equalize their influence on the similarity estimation. Finally we apply weight the normalized feature-spaces according the weights of Table 1.

4.3 Data

The data-set used within the experimental evaluation of this paper has been developed within the scope of Europeana Sounds project[23]. Out of the several hundreds of thousands of audio records accessible through Europeana APIs, a subset of more than 83.000 items forms the bases of the current evaluation. The description of these audio items was collected through the Europeana Record API[2]. Even if the representation of item description is available in a standardized knowledge graph format (using Europeana Data Model[3]), there are still several challenges to effectively use this metadata for information retrieval purposes.

The Europeana Records are collected from various Institutions from all over the Europe and pre-processed by so called national or thematic aggregators. As shown in Figure 2, the records included

Figure 3: Metadata value density / sparsity: number provided values per metadata attribute in descending order. Top ranked attributes are densely, low ranked attributes are sparsely provided.
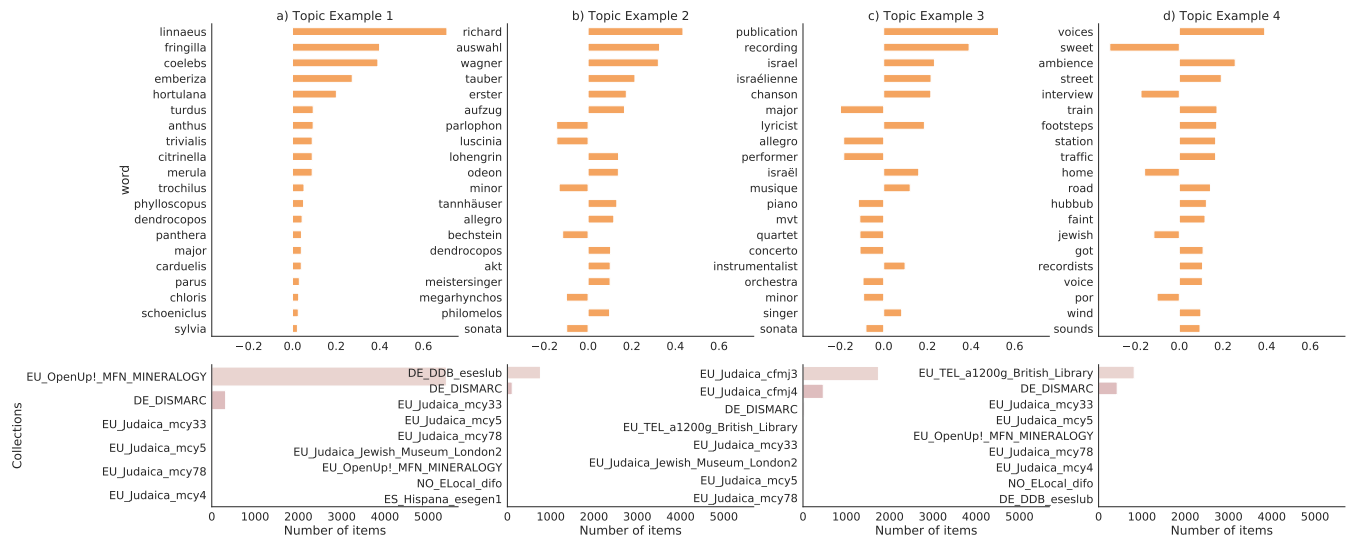
within the evaluation dataset were originally provided by 30 different institutions (also called data providers) and included in 26 different data collections. The *Europeana Sounds* project, with *Dismarc* music collection and several other non music collections, is the main thematic aggregator for sound content. It is followed by the Jewish Heritage Network, the aggregator of Judaica collections, which include many traditional songs collected from Jewish communities that were leaving in different European countries. However, many records were submitted to Europeana by national aggregators and previous research projects.

The large variety and data sparsity within this evaluation corpus represent a great challenge for the efficiency of the proposed approach. This variation is encountered with respect to the type of content, the used categorization schemes and with respect to the distribution of audio features. The largest part of sound content is represented by music records, however an important part of the data set includes radio news, public speeches or language dialects recordings, environmental (e.g. city noise) or biodiversity sounds (e.g. bird twittering), etc. In the case of music records, there is a high variety of music genres, from traditional to classical music, from love songs to rock, from instrumental to single voice singer, etc. As consequence, the evaluation dataset builds a sparse data matrix, except for the fields that are enforced as mandatory through the Europeana Data Model (see Figure 3).

Another challenge of the evaluation corpus is represented by the multilingualism of the metadata, more than 40 European languages being now used to describe the Europeana records. Even if the several data fields contain very precise keywords describing the audio

Figure 4: LSI Topic Examples visualizing the correlation between topics and collections. top - plots showing words ranked by their absolute (positive/negative) influence on the topic. bottom - number of items per collection associated with this topic. a) biological/zoological terms - mostly correlated with OpenUp! collection of animal sounds. b) terms and names of classical music and composers - mostly correlated with collection of classical music. c) terms referring to location - in this case *Israel* correlated with collection of Jewish music. d) terms describing ambient sounds - mostly correlated with a collection of environmental sounds.

content, they have limited usage for similarity search due to their language distribution. In Figure 4 we showcase the composition of the topics and indicate the influence of individual keywords. Through the topic composition, a human user can easily recognize topics relating to biodiversity and animal sounds, classical music and composers, locations, ambient sounds, etc. Europeana data collections are meant to group records that share certain properties in common, the enclosed records being similar to each other in a broader sense. However, in many cases, data collections group items from different institutions and cover several topics. The correlation between topics and data collections is indicated in Figure 4.

From the full description of audio items, 22 metadata elements were taken for computing item similarities. Some of these elements indicate the provenance and the aggregation process of audio content. The Institution that owns the content is named within the *Data Provider* element, while the *Provider* indicates the organization aggregating the content in the collection defined though the *Collection-Name* and *Collection-id*. The country and the language used by the contributing institution are also available in the metadata (i.e. *DataProvider Language*, *DataProvider Country*). The elements describing the audio content include the a mandatory *Title* and optional *Subject* categorizations, *Type* of content and textual *Description*. Quite often, within the contributor fields the role of the person is also indicated, or in case of orchestras, the music instrument might be available as well. When known, the *Date* and the *Year* are available directly in the medata elements. The approximate date and location can be indicated either through the *Coverage* or through the specialized *termsTemporal* and *termsSpatial*, *termsCreated* elements. For some items the the storage medium (*termsMedium*) and the original work from which the current object

was derived are also indicated. All these data elements contain information that has correlation with the characteristics of the audio content.

4.4 Data Pre-processing

4.4.1 Text-data Pre-processing. This section refers to the text-input to build the LSI models. All meta-data attributes are concatenated to a single string per instance using white-spaces as separators. The entire text content is converted to lower-case and HTML tags, sequences and hyperlinks are removed. Year dates are mapped to decades and centuries and translated to Roman numerals. Further, numbers and punctuation's are removed. After tokenization, stopwords are removed for the languages *English, German, Italian, French* and *Romanian*. Finally, words w with $|w| <= 2$ are also removed.

4.4.2 Audio Data Pre-Processing. All audio-files of the collection are re-sampled to 44.100 Hz and single-channel converted. An audio segment s of length $|s| = 6$ seconds is read from an audio-file a using an offset of $o = 5$ seconds to avoid silent sections at the beginning of a recording as well as fade-in effects. If $|a| <= |s| + o$ the offset is reduced to $o = |a| - |s|$. If $|a| < |s|$, the audio-file is shorter than the expected sequence length and the missing content is zero-padded. Short-time Fourier Transform (STFT) with a Hanning-windowing function and a 4096 samples window size with 50% overlap is applied. The resulting Spectrogram is transformed to the log-scaled Mel-space using 80 Mel-filters and cut-off frequencies of 16Hz (min) and 18.000Hz (max). The final shape of the DNN input matrix shape is 80x130x1. Instead of normalizing the feature-space, we add a batch-normalization layer on top of the network (see Figure 1).
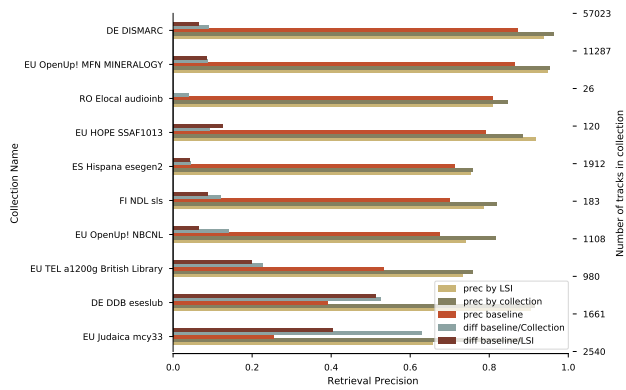
Figure 5: Evaluation results *Similarity by collection membership* measured in retrieval precision at cut-off 1 and number of tracks of the corresponding collection (right Y-axis). Global mean precision at different cut-offs presented in Table 2.
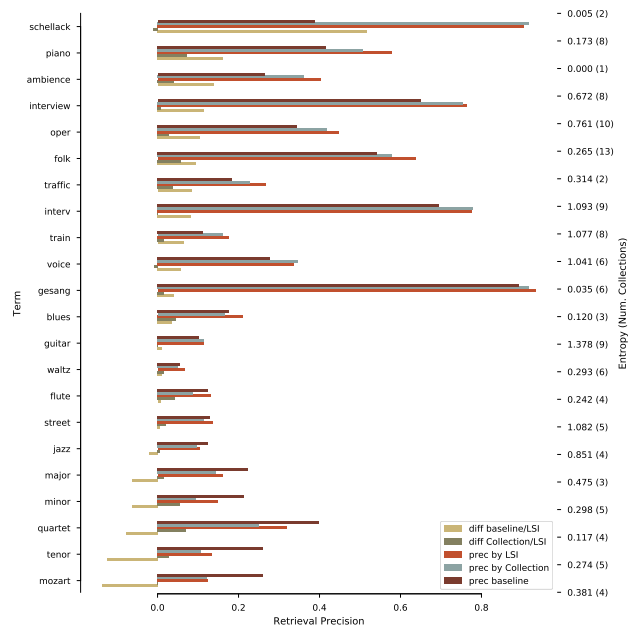

Figure 6: Evaluation results *Similarity by content related terms* measured in retrieval precision at cut-off 1. Right Y-axis shows term entropy across the number of collections. Global average precision at different cut-offs presented in Table 2.

5 RESULTS

Results are calculated identically for all experiments. First, the hand-crafted features or trained models are applied to the audio collection to embed it into a feature-space. On this, nearest-neighbor search is performed using Euclidean distance as similarity function. To compare the results and evaluate the approaches, retrieval precision at different cut-offs (1,2,3,5,10,50,100) is assessed. The model performance is then evaluated according two criteria.

(1) **Collection Similarity:** collections provided by partner libraries to the Europeana have been carefully aggregated, are coherent and share common attributes and characteristics.

(2) **Term Similarity:** content related terms such as music instruments, composers, music genres or animal species are matched in the metadata. These terms span across several collections and calculate their entropy to assess if they are evenly distributed or skewed towards a certain collection.

Track-relatedness by Collection Membership: The first evaluation is based on common approaches to representation learning which use membership to a class [8], label [24] or identity [19, 27] to select positive and negative examples for triplet based neural networks. As can be observed in Figure 5 this approach tends to learn representation which focus on collection related features and acoustic artifacts. From this perspective, this approach seems to outperform the baseline as well as the proposed approach, but, although, the collections of the dataset have been well aggregated by professional librarians, the learned representation do not generalize towards to a global acoustic similarity, as can be observed in Figure 6. The term-based precision values for the model trained on track-relatedness by collection membership differ recognizably from the LSI based approach, especially for evaluation terms with high entropy values.

Track-relatedness by LSI-Topic Similarity: The potential of the proposed LSI-based representation learning approach is shown

by the decrease of precision in the collection-based and the increase in the term-based evaluation. This indicates that the learned representation capture characteristics which facilitate acoustic similarity estimations globally, across the collections of the dataset. For some terms the representations learned from free text outperform the baseline significantly, such as for the acoustic characteristics of digitized shellac recordings which are difficult to model with handcrafted audio features. It can generally be observed that the representation learned by our approach improves over the baseline in terms of capturing general acoustic characteristics such as ambient sounds, sound producing objects such as instruments or machines as well as human voice. Regarding the first baseline following the feature-based approach presented in [25], our proposed approach does not capture music related characteristics accordingly. An explanation for this could be, because only 6 dimensions of the feature-set of the baseline approach are general audio features wheres the remaining 1671 dimensions belong to music features. Regarding the second baseline using collection membership as track-relatedness measure, the LSI-based track-relatedness approach improved over almost all term-based results.

Discussion: Audio similarity is generally difficult to evaluate. Usually, categorical ground truth assignments are used alternatively and precision is defined on retrieving tracks belonging to the same category [13]. In this paper we describe an approach to learn an audio representation for similarity estimation from multi-lingual free-text under the absence of ground-truth data. Thus, we approach the evaluation from two perspectives: the categorical view

Table 2: Overview of baseline audio representation. Weighted composition of state-of-the-art audio and music content *Features*. The audio-content *descriptors*, their corresponding acoustic *categories*, their assigned feature *weight* as well as the cumulative *category weight*.

	Dim	1	2	3	5	10	50	100
baseline	1677	**0.288**	**0.253**	**0.235**	0.213	0.188	0.139	0.121
by Collection	256	0.232	0.216	0.208	0.196	0.182	0.156	0.143
by LSI	256	0.264	0.241	0.231	**0.218**	**0.201**	**0.170**	**0.158**

using collection memberships and terms in the metadata describing acoustic properties. With this approach we are able to show that the LSI based approach to audio representation learning provides similar retrieval precision to the feature-based baseline approach up to a cut-off of 3 (see Table 2). From a cut-off of 5 our learned representations with 256 dimensions exceed the baseline with 1677 dimensions. Thus, the size of the feature-space is reduced by a factor of 6.5 at consistent performance.

6 CONCLUSIONS AND FUTURE WORK

We introduced a novel approach to unsupervised audio representation learning. We showed how to estimate track-relatedness from unstructured multilingual free-text by projecting the semantic data into a vector space using Latent Semantic Indexing. This track-relatedness is used for online triplet selection to train a triplet deep neural network which cross-learns an audio representation from the text modality. We showed that the representations learned perform similar to the baseline, respectively exceeding the baseline in similarity retrieval precision at higher cut-offs at only 15% of the baseline's feature vector length.

REFERENCES

[1] Adam Berenzweig, Beth Logan, Daniel PW Ellis, and Brian Whitman. 2004. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal* 28, 2 (2004), 63–76.
[2] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR2011)*, Vol. 2. 10.
[3] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. 2017. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2392–2396.
[4] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41 (1990), 391–407.
[5] Simon Dixon. 2007. Evaluation of the audio beat tracking system beatroot. *Journal of New Music Research* (2007).
[6] Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D Plumbley. 2013. Detection and classification of acoustic scenes and events: An IEEE AASP challenge. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 1–4.
[7] Christopher Harte, Mark Sandler, and Martin Gasser. 2006. Detecting harmonic change in musical audio. In *Proc. 1st ACM WS on Audio and music computing multimedia*.
[8] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*. Springer, 84–92.
[9] Eric J Humphrey, Juan P Bello, and Yann LeCun. 2013. Feature learning and deep architectures: New directions for music informatics. *Journal of Intelligent Information Systems* 41, 3 (2013), 461–481.

[10] Peter Knees and Markus Schedl. 2013. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2013).
[11] Peter Knees and Markus Schedl. 2015. Music retrieval and recommendation: A tutorial overview. In *Proc. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*.
[12] Peter Knees and Markus Schedl. 2016. Contextual Music Similarity, Indexing, and Retrieval. In *Music Similarity and Retrieval: An Introduction to Audio- and Web-based Strategies*. Springer Berlin Heidelberg, Berlin, Heidelberg, 133–158. https://doi.org/10.1007/978-3-662-49722-7_6
[13] Peter Knees and Markus Schedl. 2016. *Music similarity and retrieval: An introduction to audio-and web-based strategies*. Vol. 36. Springer.
[14] Thomas Lidy and Andreas Rauber. 2005. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *ISMIR*.
[15] Thomas Lidy, Andreas Rauber, Antonio Pertusa, and José Manuel Iñesta Quereda. 2007. Improving Genre Classification by Combination of Audio and Symbolic Descriptors Using a Transcription Systems.. In *Proc. Int. Conf. Music Information Retrieval*.
[16] Thomas Lidy and Alexander Schindler. 2016. CQT-based Convolutional Neural Networks for Audio Scene Classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. 60–64.
[17] Beth Logan and Ariel Salomon. 2001. A Music Similarity Function Based on Signal Analysis.. In *ICME*. 22–25.
[18] Meinard Müller. 2015. *Fundamentals of music processing: Audio, analysis, algorithms, applications.* Springer.
[19] Jiyoung Park, Jongpil Lee, Jangyeon Park, Jung-Woo Ha, and Juhan Nam. 2018. Representation learning of music using artist labels. In *19th International Society for Music Information Retrieval Conference (ISMIR 2018)*.
[20] Yves Raimond, Samer A Abdallah, Mark B Sandler, and Frederick Giasson. 2007. The Music Ontology.. In *ISMIR*, Vol. 2007. Citeseer, 8th.
[21] Ubai Sandouk and Ke Chen. 2016. Learning Contextualized Music Semantics from Tags Via a Siamese Neural Network. *ACM Transactions on Intelligent Systems and Technology* 8, 2 (2016), 1–21. https://doi.org/10.1145/2953886
[22] Alexander Schindler, Martin Boyer, Andrew Lindley, David Schreiber, and Thomas Philipp. 2019. Large Scale Audio-Visual Video Analytics Platform for Forensic Investigations of Terroristic Attacks. In *International Conference on Multimedia Modeling*. Springer, 106–119.
[23] Alexander Schindler, Sergiu Gordea, and Harry van Biessum. 2016. The europeana sounds music information retrieval pilot. In *Euro-Mediterranean Conference*. Springer, 109–117.
[24] Alexander Schindler and Peter Knees. 2019. Multi-Task Music Representation Learning from Multi-Label Embeddings. In *Proceedings of the 17th International Workshop on Content-based Multimedia Indexing (CBMI 2019)*. Dublin, Ireland.
[25] Alexander Schindler, Thomas Lidy, and Andreas Rauber. 2016. Comparing shallow versus deep neural network architectures for automatic music genre classification. In *9th Forum Media Technology (FMT2016)*, Vol. 1734. 17–21.
[26] Alexander Schindler, Rudolf Mayer, and Andreas Rauber. 2012. Facilitating comprehensive benchmarking experiments on the million song dataset. In *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR 2012)*.
[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
[28] Siddharth Sigtia and Simon Dixon. 2014. Improved music feature learning with deep neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 6959–6963.
[29] George Tzanetakis and Perry Cook. 2000. Marsyas: A framework for audio analysis. *Organised sound* 4, 3 (2000), 169–175.
[30] Henning Scholz Walter Koch. 2009. DISMARC and BHL-Europe: multilingual access to two aggregation platforms for Europeana. In *Proceedings of the Workshop on Advanced Technologies for Digital Libraries 2009*. Trento, Italy, 25–29.
[31] Yi-Hsuan Yang and Homer H Chen. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 3 (2012), 40.