AUDIETO-BASED DISTRIBUTIONAL SEMANTIC MODELS
FOR MUSIC AUTO-TAGGING AND SIMILARITY MEASUREMENT

Giannis Karamanolakis1  Elias Iosif1  Athanasia Zlatintsi1
Aggelos Pikrakis2  Alexandros Potamianos1

1School of Electrical & Computer Engineering, National Technical University of Athens, Greece
2Department of Informatics, University of Piraeus, Greece

giannis.karamanolakis@gmail.com, iosife@central.ntua.gr, nzlat@cs.ntua.gr,
pikrakis@unipi.gr, potam@central.ntua.gr

ABSTRACT

The recent development of Audio-based Distributional Semantic Models (ADSMs) enables the computation of audio and lexical vector representations in a joint acoustic-semantic space. In this work, these joint representations are applied to the problem of automatic tag generation. The predicted tags together with their corresponding acoustic representation are exploited for the construction of acoustic-semantic clip embeddings. The proposed algorithms are evaluated on the task of similarity measurement between music clips. Acoustic-semantic models are shown to outperform the state-of-the-art for this task and produce high quality tags for audio/music clips.

Index Terms— distributional semantic models, bag-of-audio-words, auto-tagging, music similarity

1. INTRODUCTION

Semantic information in the form of metadata, e.g., tags, has been valuable in enhancing the performance for many music processing tasks [1, 2, 3]. Metadata typically comes in two forms: free text associated with a music or audio clip or tags (list of words or phrases) that describe the clip. Tags are often preferred over (web-mined) text, because they give a direct description of the song, e.g., genre or instruments, while the latter is inherently noisy; only a part of the text is musically relevant. The automatic annotation of clips (auto-tagging) is becoming vital and finds numerous applications including efficient music indexing and retrieval.

A variety of auto-tagging methods have been proposed in the literature. In [4], the similarity between artists is exploited for the prediction of the most descriptive tags for clips. In [5], language models were computed using Restricted Boltzmann Machines, while in [6], the combination of audio features is proposed within a block-level framework. The use of semantic tags for music similarity measurement is proposed in [7], where each song is represented by a Semantic Multinomial Distribution over a vocabulary of tags.

Music similarity is at the core of query-by-example, where the user gives a musical piece as a query and the system returns a ranked list of recommendations. Content-based similarity can be exploited by collaborative filtering algorithms [8] especially where there is lack of collaborative filtering data, a.k.a. “cold start” problem. This improves the efficiency of music recommendation and playlist generation, two important tasks in Music Information Retrieval (MIR). Music similarity estimation can be formulated as the problem of finding an appropriate embedding of a music clip with respect to a distance metric. Many approaches use machine learning techniques in order to learn the distance metric that best approximates absolute user ratings [9, 10] or relative user ratings from different datasets [11]. Furthermore, research efforts include the collection of similarity scores that can be used as groundtruth information [12, 13] and the investigation of different evaluation techniques [14] for music similarity.

Distributional Semantic Models (DSMs) [15] is a popular method for automatically constructing semantic representations from text. Despite their success in various semantic tasks (e.g., semantic classification and computation of semantic similarity), the DSMs have been criticized as “disembodied”, since they rely solely on linguistic information without being grounded in perception and action. The disconnection of natural language from the physical world, also referred as the symbol grounding problem [16], is alleviated via the integration of multiple modalities [17, 18]. The development of audio-based DSMs (ADSMs) was proposed in [19] for the representation of words based on their acoustic properties, while in [20] an extension was presented using combinations of auditory and linguistic features. A recently proposed approach dealt with the fusion of (different) acoustic features according to the nature of sounds (music, speech, other) [21].

In this work, the ADSM described in [21] is adopted for the computation of audio and lexical vector representations in a joint acoustic-semantic space. These ‘bag-of-audio-words’ representations are used for the automatic annotation of music clips. Then, the predicted tags and the acoustic features are exploited for the construction of acoustic-semantic clip embeddings. The proposed algorithms are evaluated on the task of music similarity measurement between clips taken from the MagnaTagATune dataset [12].

This paper is organized as follows. In Section 2, various methods are described for the construction of clip representations. In Section 3.1, the experimental dataset and procedure are described in detail, while in Section 4, the evaluation results of the proposed methods are reported and compared with the literature.

2. SYSTEM DESCRIPTION

For a dataset consisting of audio clips and the corresponding annotation (tags), the procedure for the creation of clip representations consists of a series of steps described below.
2.1. Creation of audio-word vocabulary

In order to train an audio-word vocabulary, the clips are partitioned into partially overlapping windows and a feature vector is extracted from every window. Therefore, every clip is represented by a set of vectors depending on its length. Next, all vectors are clustered by applying the $k$-means algorithm and the $k$ centroids of the returned clusters are considered as the audio-words of the vocabulary.

2.2. Audio clip representations

As presented in [21], audio clips are represented as a mixture of audio-words. For each window $\mathbf{x}_t$ of a clip, a feature vector $\mathbf{x}_t \in \mathbb{R}^d$ is computed (where $d$ is the dimensionality of the feature space) and associated with the audio-word vocabulary. The association is performed by encoding the $\mathbf{x}_t$ vector as a $k$-dimensional vector $\mathbf{e}_t$:

$$\mathbf{e}_t = (w_1, w_2, ..., w_k),$$

where $w_i \in [0, 1]$ and $\sum_{i=1}^{k} w_i = 1$. The weight $w_i$ denotes the contribution of the $i$-th audio-word to the window representation and is computed according to the similarity score between the audio-word and the feature vector $\mathbf{x}_t$ (see [21] for details). Finally, the clip representation, $r_c$, is obtained by averaging the vectors computed for the respective windows. Given a collection consisting of $M$ clips, this process results in a $M \times k$ matrix. The space consisting of bag-of-audio-words clip representations, will be referred as AUDIO space.

2.3. Tag representations via the ADSM

As described in [21], an ADSM is constructed that provides bag-of-audio-words embeddings for tags, based on their association with the clips. In particular, the representation $r_j$ of a tag $j$ is computed by averaging the representations of clips that have this tag in their descriptions. For a collection of audio clips with $T$ (unique) tags this results in a $T \times k$ matrix. Then, the Positive Pointwise Mutual Information (PPMI) weighting is applied to the matrix [22]. The overall procedure, up to the creation of tag representations, is represented in Fig 1(a).

2.4. Clip annotation (auto-tagging)

The bag-of-audio-words representations of tags provide a straightforward way for the automatic annotation (auto-tagging) of an audio clip $c$. First, the clip gets a representation, $r_c$, in the AUDIO Space. Then, the cosine similarity is computed between $r_c$ and the representation $r_j$ of each tag $j$:

$$s_{cj} = \frac{r_c \cdot r_j}{|r_c| \cdot |r_j|}.$$  (2)

The $N$ tags that best describe the clip $c$ are those corresponding to the $N$ highest similarity scores $s_{cj}$.

Table 1 includes some examples selected from the MagnaTagATune dataset (see Section 3.1) for which the groundtruth labels are compared with the $N = 5$ automatically predicted tags. It is observed that many tags appear both in the predicted and the groundtruth labels, while other tags have very similar meaning (e.g., ‘silence’ and ‘quiet’). Moreover, quite descriptive tags are returned for clips that have no annotations (e.g., clip 19920).

2.5. Semantic representations of clips

The ADSM described in Section 2.3 is exploited for the representation of audio clips via their association with tags. Specifically, the embedding of a clip is obtained by averaging the representations of tags that are semantically related with the clip. The determination of clip-tag associations can be performed either using already provided metadata, or via the proposed auto-tagging scheme. In the latter case, the clip gets a representation (in the ADSM-AUTOTAG space), as follows:

$$r_c' = \frac{1}{N} \sum_{i=1}^{N} r_i,$$  (3)

where $r_i$ is the representation of the $i$-th tag, while $N$ denotes the number of tags returned by the auto-tagger.

2.6. Fusion of audio and semantic representations

The similarity score between two clips in the AUDIO space is computed as the cosine similarity of their acoustic representations. Sim-
Due to the objective opinion of each user, there are constraints that contradict each other. Hence, a method was proposed in [10] in order to deal with the inconsistent constraints. In [9], the 860 remaining constraints derived from [10] were split into non-overlapping training and test sets of 774 and 86 constraints, respectively enabling 10-fold cross-validation. These constraints are published serving as a common evaluation benchmark. The described benchmark is used in this work.

3.2. Data preprocessing and feature extraction

All audio clips are converted to WAV format and resampled at 22.05 kHz. For each clip, a feature vector is extracted from windows of 250 ms with a step of 100 ms. Here, two different feature vectors are extracted: the first (EchoNest) is obtained by the EchoNest API 1.0\(^2\). Specifically, 24 features per audio frame have been kept, consisting of 12 chroma features and 12 timbre features. The chroma features describe the relative dominance of every pitch in the chromatic scale and are normalized to [0, 1]. The timbre features correspond to the coefficients of 12 basis functions which represent the texture of sound. The second type of feature vectors (MFCCd) consists of the Mel Frequency Cepstral Coefficients (MFCCs) (concatenated with spectral energy), their 1st and 2nd order derivatives, resulting in a vector of 39 coefficients. In both cases, the feature vectors are normalized by their mean and standard deviation values (Z-normalization).

3.3. Experimental procedure

For each step of the 10-fold cross validation, the audio-word vocabulary and the ADSM are built using the training clips\(^3\). The representations are then computed for the test clips and the similarity scores are obtained, as described in Section 2.6. The following methods are evaluated:

- **AUDIO**: only the acoustic features are used (Section 2.3).
- **ADSM**: clip representations are obtained via the provided tags (Section 2.5).
- **ADSM-AUTOTAG**: same with ADSM method, but all clip representations (even for clips that have labels) are derived via auto-tagging (Section 2.4). Here, \(N = 20\) tags are predicted for every clip.

\(^2\)http://developer.echonest.com/

\(^3\)For computational efficiency, 1000 clips are randomly selected from the training set for the audio-word vocabulary.

### Table 1: Examples of auto-tagging outputs for the MagnaTagATune dataset.

| Clip id | Groundtruth Tags | Predicted Tags (N=5) |
|---------|------------------|---------------------|
| 3843    | *indian, sitar*  | sitar, *indian*, eastern, india, oriental |
| 9531    | *rock, heavy, heavy metal, loud, fast, hard rock, metal* | hard, loud, *heavy* metal, metal |
| 13526   | *bass, drums, drum, funky, reggae* | funky, beat, drums, reggae, funk |
| 15380   | *classical, solo, cello, violin, strings* | cello, *viola*, *violin*, solo, classical |
| 19920   | -                | orchestra, violins, flutes, fiddle, violin |
| 21725   | *choir, choral, men, man* | monks, chant, chanting, *men*, choral |
| 29231   | *acoustic, guitar* | classical guitar, *guitar*, acoustic, lute, spanish |
| 43390   | *rock, loud, pop, vocals, male vocals* | *male vocals*, pop, male vocal, male singer, *rock* |
| 48010   | silence          | low, soft, no singing, quiet, wind |
| 57081   | *piano*          | piano solo, *piano*, classic, solo, classical |

1. http://magnatune.com/

Similarly, the similarity between clips in the ADSM space is computed as the cosine similarity of their respective semantic representations. However, the characterization of clips only via their semantic representations in some cases may lead to inaccurate estimates. For example, if two clips are annotated with the same labels, they will have exactly the same ADSM representations although they sound different. This problem can be alleviated with the fusion of acoustic and semantic information. The clip embedding \(r_c\) in the AUDIO space can be combined with the embedding from the ADSM (or ADSM-AUTOTAG) space, \(\vec{r}_c\), via a weighted average scheme:

\[
\vec{r}_c = w \vec{r}_c + (1 - w) \vec{r}_c,  
\]

or via a weighted concatenation:

\[
\vec{r}_c = w \vec{r}_c \oplus (1 - w) \vec{r}_c,  
\]

where \(\oplus\) stands for the vector-concatenate operator and \(w\) is a real-valued number indicating the relative importance of the fused representations. This method results in clip embeddings in the FUSION (or FUSION-AUTOTAG) space. The various methods for the representation of clips are graphically represented along with an auto-tagging example in Fig 1(b).

### 3. EXPERIMENTAL DATASET AND PROCEDURE

The proposed algorithms are evaluated for the similarity computation between songs. For this purpose, the MagnaTagATune dataset is used.

#### 3.1. The MagnaTagATune dataset

The MagnaTagATune dataset contains 25,863 30-second audio clips (provided by the Magnatune\(^1\) label) and 188 tags. In addition, similarity data have been collected from the TagATune game [23]. In a bonus part of the game, the user listens to three songs and gives a vote to the song that sounds as the most dissimilar when compared with the other two songs (often called as odd one out game). The similarity data for every triplet of clips is stored in form of a vote to the song that sounds as the most dissimilar when compared with the other two songs (often called as odd one out game).

The similarity data for every triplet of clips is stored in form of a vote to the song that sounds as the most dissimilar when compared with the other two songs (often called as odd one out game). For this purpose, the MagnaTagATune dataset is used. Due to the objective opinion of each user, there are constraints that contradict each other. Hence, a method was proposed in [10] in order to deal with the inconsistent constraints. In [9], the 860 remaining constraints derived from [10] were split into non-overlapping training and test sets of 774 and 86 constraints, respectively enabling 10-fold cross-validation. These constraints are published serving as a common evaluation benchmark. The described benchmark is used in this work.

1. http://magnatune.com/
In this work, an audio-based DSM is exploited for creating lexical representations and the clip representations are derived via an auto-tagging scheme. The evaluation is performed on the task of music similarity measurement using the MagnaTagATune dataset. Although the similarity ratings provided with the dataset are not considered in the proposed algorithms, they exceed state-of-the-art methods, that use them in order to train a similarity metric. Moreover, very few data, e.g., 50 clips, are adequate to train the vocabulary, which is used for the creation of bag-of-audio-words representations. Therefore, this unsupervised algorithm can be applied in every dataset where neither clip annotations or similarity data are necessary. Regarding future work, we aim to investigate the fusion of acoustic, semantic and visual modalities and test the proposed algorithm on multimedia data.

6. ACKNOWLEDGEMENTS

This work has been partially supported by the BabyRobot project supported by EU H2020 (grant # 687831).
7. REFERENCES

[1] Mark Levy and Mark Sandler, “Music information retrieval using social tags and audio,” IEEE Transactions on Multimedia, vol. 11, no. 3, pp. 383–395, 2009.

[2] Yu-Ching Lin, Yi-Hsuan Yang, and Homer H Chen, “Exploiting online music tags for music emotion classification,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 7, no. 1, pp. 26, 2011.

[3] Paul Lamere, “Social tagging and music information retrieval,” Journal of New Music Research, vol. 37, no. 2, pp. 101–114, 2008.

[4] Joon Hee Kim, Brian Tomasik, and Douglas Turnbull, “Using artist similarity to propagate semantic information,” in International Society of Music Information Retrieval, 2009, vol. 9, pp. 375–380.

[5] Michael I Mandel, Razvan Pascanu, Douglas Eck, Yoshua Bengio, Luca M Aiello, Rossano Schifanella, and Filippo Menczer, “Contextual tag inference,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 7, no. 1, pp. 32, 2011.

[6] Klaus Seyerlehner, Markus Schedl, Peter Knees, and Reinhard Sonnleitner, “Draft: A refined block-level feature set for classification, similarity and tag prediction,” Dept. of Computational Perception, Johannes Kepler University, Linz, Austria, 2011.

[7] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet, “Semantic annotation and retrieval of music and sound effects,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 2, pp. 467–476, 2008.

[8] Brian McFee, Luke Barrington, and Gert Lanckriet, “Learning content similarity for music recommendation,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2207–2218, 2012.

[9] Daniel Wolff, Sebastian Stober, Andreas Nürnberger, and Tillman Weyde, “A systematic comparison of music similarity adaptation approaches,” in International Society of Music Information Retrieval, 2012, pp. 103–108.

[10] Sebastian Stober and Andreas Nürnberger, “An experimental comparison of similarity adaptation approaches,” in International Workshop on Adaptive Multimedia Retrieval, 2011, pp. 96–113.

[11] Daniel Wolff, Andrew MacFarlane, and Tillman Weyde, “Comparative music similarity modelling using transfer learning across user groups,” Proceedings of the 16th ISMIR Conference, 2015.

[12] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie, “Evaluation of algorithms using games: The case of music tagging,” in International Society of Music Information Retrieval, 2009, pp. 387–392.

[13] Daniel PW Ellis, Brian Whitman, Adam Berenzweig, and Steve Lawrence, “The quest for ground truth in musical artist similarity,” in International Society of Music Information Retrieval. Paris, France, 2002.

[14] Beth Logan, Daniel PW Ellis, and Adam Berenzweig, “Toward evaluation techniques for music similarity,” The MIR/MDL Evaluation Project White Paper Collection, vol. 3, pp. 81–85, 2003.