A DIMENSIONAL CONTEXTUAL SEMANTIC MODEL FOR MUSIC DESCRIPTION AND RETRIEVAL

Michele Buccoli, Alessandro Gallo, Massimiliano Zanoni, Augusto Sarti, Stefano Tubaro

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Piazza Leonardo da Vinci 32 - 20133 Milano, Italy
{michele.buccoli, massimiliano.zanoni, augusto.sarti, stefano.tubaro}@polimi.it

ABSTRACT

Several paradigms for high-level music descriptions have been proposed to develop effective system for browsing and retrieving musical content in large repositories. Such paradigms are based on either categorical or dimensional models. The interest in dimensional models has recently grown a great deal, as they define a semantic relation between concepts through graded descriptions. One problem that affects semantic descriptions is the ambiguity that often arises from using the same descriptor in different contexts. In order to overcome this difficulty, it is important to model and address polysemy, which is the property of words to take on different meanings depending on the use-context. In this paper we propose a Dimensional Contextual Semantic Model for defining semantic relations among descriptors in a context-aware fashion. This model is here used for developing a semantic music search engine. In order to evaluate the effectiveness of our model, we compare this engine with two systems that are based on different description models.

Index Terms— music, information retrieval, polysemy, semantic contexts, music recommendation

1. INTRODUCTION

In the past few decades, the importance of high-level (semantic) description of musical content has progressively grown to the point of becoming a fundamental task in music retrieval applications. In particular, Query By Semantic Description (QBSD) [1, 3, 4] based on dimensional approaches has gained a great deal of popularity. In dimensional approaches, terms are represented in a metric space where the distance describes semantic similarity. One of the most widely used model [5, 6, 7] for emotion-related music description is the Valence-Arousal (VA) space [8], where words are expressed in terms of degree of positivity (Valence) and of activation (Arousal). A second well-known approach takes advantage of user tagging [9, 10] and infers the semantic relation between terms by exploiting the co-occurrence of descriptors [14, 11, 12]. A semantic space is built through Latent Semantic Indexing (LSI) [13], which assumes that the semantic correlation among descriptors is proportional to the frequency of their joint use for annotation. Such methods proved quite effective, though they tend to represent all the terms in a single semantic space. This does not account for the fact that the meaning of terms in natural language could change depending on the context (*polysemy*). For example, the term *soft* could be describe timbral properties, but also emotional proprieties; and the term *quiet* could refer to the mood as well as the performance style.

Ignoring polysemy introduces bias in music description. For example, when using LSI to infer semantic relations, terms such as *Anxious* and *Hard* end up very close to each other, as they are frequently used together. Also in the VA space such terms end up near each other, and far from the term *Soft*. This happens because in music description such as *Hard* and *Soft* are often used for addressing timbral proprieties. In other contexts, however, this could be misleading, as *Anxious* songs could be thought of as *Hard* as well as *Soft*. Approaches based on non-negative matrix factorization (NMF) have been proven to be effective for solving polysemy issues [17]. However, they still have not been applied to scenarios where polysemy involves nuances of meaning in the same broad topic (e.g. music description).

In this study, we propose a music description paradigm based on a Dimensional Contextual Semantic Model (DCSM). High-level descriptors are grouped into contexts that represent different aspects of music and each term can belong to several contexts, in order to account for polysemy. Within the contexts, the semantic relationship between descriptors is modeled by means of a graded semantic similarity that ranges from antonymy, to neutrality, to synonymy. This approach can address the bias in music description and the misleading interpretation of the user query by exploiting semantic similarities and context information. *A-priori* information about contexts and similarity among descriptors have been collected though surveys. In order to evaluate the effectiveness of the approach, we developed a prototype of a search engine based on DCSM that deals with natural language queries and

This research activity has been partially funded by the Cultural District of the province of Cremona, Italy, a *Fondazione CARIPLO* project, and by the Arvedi-Buschini Foundation

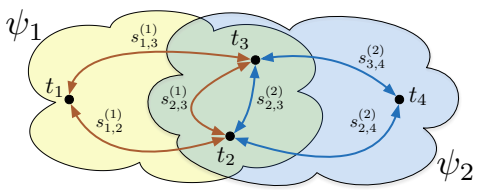

Fig. 1: A representation of the DCSM with the terms $t_1, ...t_4$ and the overlapping contexts ψ_1, ψ_2, where their semantic relationship is modeled by $s_{ij}^{(k)}$ with $i, j = 1, ..., 4$ and $k = 1, 2$.

we compared it with a system based on the well-known LSI paradigm and with the system proposed in [4].

2. CONTEXTUAL SEMANTIC SPACES

We define the DCSM and how it includes the effect of polysemy. We then examine how the musical items and user queries can exploit the DCSM to overcome annotation or query ambiguity.

2.1. Semantic model

Given a vocabulary $\mathcal{V} = \{t_1, ..., t_M\}$ of M terms, we define a *context* ψ_k as a subset of $\mathcal{V}$ that represents a specific musical aspect. A descriptor t_i is in the context ψ_k if it *has a meaning* within that context. In order to model the polysemy, the sets of context can generally have an overlap, therefore a term may belong to multiple contexts. As an example, Fig. 1 depicts a possible scenario with two contexts ψ_1 and ψ_2 and 4 terms t_1, t_2, t_3, t_4. The terms t_2, t_3 belong to both the contexts.

We model the semantic relationship between terms by assigning a *similarity score* $s_{ij}^{(k)}$ to each pair of descriptors t_i, t_j in each context ψ_k. More formally:

$$s_{ij}^{(k)} = s_{ji}^{(k)} \in [-1, 1] \; \forall \; t_i, \; t_j \in \mathcal{V}, \; \forall \; k = 1, ..., K, \quad (1)$$

where K is the number of contexts. A negative value of $s_{ij}^{(k)}$ represents a degree of antonymy, whereas a positive value represents a degree of synonymy in the context ψ_k. The 0 value expresses the absence of a semantic relation exists between the two terms in the context ψ_k. This is also the case of one of the terms that does not have a semantics in ψ_k:

$$t_i \notin \psi_k \implies s_{ij}^{(k)} = s_{ji}^{(k)} = 0 \; \forall \; j = 1, ..., M. \quad (2)$$

In the example in Fig. 1, t_2 and t_3 have (generally different) similarity scores $s_{2,3}^{(1)}$ and $s_{2,3}^{(2)}$, with respect to the two possible context ψ_1 and ψ_2 respectively. Instead, the terms t_1 and t_4 are not prone to potential ambiguities, since they have a semantic in a unique context: $s_{1,4}^{(1)} = s_{1,4}^{(2)} = 0$.

We collect a set of K symmetric *similarity matrices* $\mathbf{S}^k \in \mathbb{R}^{M \times M}$, which are composed by the elements $s_{ij}^{(k)}$. Such matrices will be used to enrich a music description and to disambiguate terms that exhibit polysemy in a query from user.

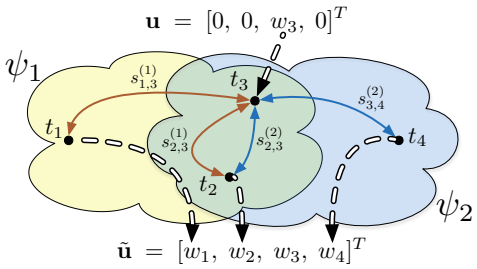

Fig. 2: A representation of the semantic enrichment of a generic vector $\mathbf{u}$ with the DCSM by inferring the missing weigths $w_j = 0$ from $w_3 \neq 0$ and $s_{3,j}^{(k)}$.

2.2. Music Description and Query Model

Given a set $d_1, ..., d_N$ of N musical items, we describe each one as a vector $\mathbf{d_i} = [w_1, ..., w_j, ..., w_T]^T \; i = 1, ..., N$, where $w_j \in [-1, 1]$ expresses the relevance of the correspondent term t_j to describe the musical item d_i. Negative values of w_j express that the item d_i could be described with an antonym of the term t_j.

Musical items are retrieved by means of a query q that we model as a vector $\mathbf{q} = [w_1, ..., w_j, ..., w_T]^T$ with

$$w_j = \begin{cases} \rho_j & \text{if } t_j \in q \\ 0 & \text{else} \end{cases}, \quad (3)$$

where $\rho_j \in [-1, 1]$ is the desired intensity for the descriptor t_j. Negative values of ρ_j can be used to express how much a descriptor is *not* desired to be present in the retrieved items, whereas positive values represent how much it is. The 0 value is the neutral weight and expresses that the correspondent term is not relevant for the query.

Music descriptions and queries may both exhibit missing weights, i.e., $w_j = 0$. In fact, classical approaches to music pieces annotation are based on manual annotation by users (e.g., social tagging [9]) or by automatic annotation (*autotagging* [3]). This has the effect to produce description vectors $\mathbf{d_i}$ that may be weakly annotated, i.e., be annotated with only a subset of the terms in the vocabulary [9, 18]. Queries may suffer from the same issue since users are prone to use only few terms relevant for the specific request.

Ambiguity is a further issue in music description and retrieval systems. A set of tags and a query, in fact, may use ambiguous descriptors that belong to different contexts. We exploit the DCSM in order to produce full-labeled and not ambiguous description vectors $\mathbf{d}$ and queries $\mathbf{q}$ by an enrichment procedure.

2.3. Exploiting the DCSM

In the following, we will use the notation $\mathbf{u} \in \mathbb{R}^M$ to indicate a generic vector of weights. We aim at inferring the missing weights ($w_j = 0$) by means of the similarity scores $\mathbf{s}_{ij}^{(k)}$ and $w_i \neq 0$, in order to obtain an enriched vector $\tilde{\mathbf{u}}$. An intuitive

Vocabulary $\mathcal{V}$
Perceived Emotion
Aggressive, Angry, Annoyed, Anxious, Boring, Carefree, Calm, Cheerful, Depressed, Dark, Exciting, Fun, Frustrated, Funny, Happy, Joyful, Light, Nervous, Relaxed, Quiet, Sad, Serious, Sweet, Tender, Tense
Timbral Description
Bright, Clean, Dark, Hard, Harsh, Heavy, Rough, Smooth, Soft, Warm
Dynamicity
Dynamic, Calm, Fast, Flowing, Quiet, Relaxed, Slow, Static, Stuttering

Table 1: List of terms for each context cluster, obtained with the survey

representation of this procedure is shown in Fig. 2, where the missing weights $w_1 = w_2 = w_4 = 0$ are inferred by combining the present weight w_3 with the similarity scores for each context. Since t_2 and t_3 have two contexts in common, it is first needed to disambiguate the context. We address this ambiguity issue by weighting the contribution of the contexts ψ_1 and ψ_2 to $\mathbf{u}$.

We first define $\mathcal{D}^{(k)} = \{t_j \in \psi_k : w_j \neq 0\}$ as the set of terms t_j that are in the context ψ_k whose weights are defined in $\mathbf{u}$. Afterwards, we define $p(\psi_k|t_j)$ as the probability that a term t_j is used as belonging to the context ψ_k. This probability is derived by the manual annotations of the terms, as described in Section 3.1. We compute the contribution of the context ψ_k to $\mathbf{u}$

$$p(\psi_k|\mathbf{u}) = \frac{\sum_{t_j \in \mathcal{D}^{(k)}} p(\psi_k|t_j)}{\sum_{k=1}^{K} \left(\sum_{t_j \in \mathcal{D}^{(k)}} p(\psi_k|t_j) \right)}, \tag{4}$$

as the normalized sum of the contributions of the context ψ_k to the terms t_j that are present in the annotation or query.

Finally, we derive the enriched vector $\tilde{\mathbf{u}}$ by weighting the sum of the contributions of the contexts to the vector:

$$\tilde{\mathbf{u}} = \sum_{k=1}^{K} p(\psi_k|\mathbf{u}) \mathbf{S}^{(k)} \mathbf{u}. \tag{5}$$

3. MUSIC SEARCH ENGINE IMPLEMENTATION

3.1. Semantic Description

In order to test the DCSM, we compose the vocabulary $\mathcal{V}$ of $M = 40$ representative terms frequently used in music description applications (Table 1). $\mathcal{V}$ is a subset of the vocabulary proposed in the ANEW project [19] and of the vocabulary used in [20]. We selected $K = 3$ contexts that capture the following musical aspects: **Perceived Emotion** concerns the perceived mood of a song; **Timbre** refers to the sound characteristics of music; **Dynamicity** is related to the dynamic characteristics (intensity and velocity) on the music piece.

Semantic similarity scores $\mathbf{s}_{ij}^{(k)}$			
Context	Calm - Quiet	Calm-Relaxed	Quiet-Relaxed
Perceived Emotions	0.933	0.8	0.675
Dynamicity	0.833	0.867	0.3

Table 2: Similarity scores between terms *Calm*, *Quiet*, *Relaxed* in the *Perceived Emotion* and *Dinamicity* contexts

The definition of the semantics of terms and of the semantic similarity between terms is a popular problem in literature [13]. Dealing with thousand of terms makes a human annotation not affordable. In this study, thanks to the reduced vocabulary, and in order to be independent from specific solution, we collected semantic information through a two-stages survey in English language, which was available online from January 15th to February 16th, 2014.

In the first stage of the survey, a subset of randomly chosen terms was proposed to each tester, who was asked to assign each term to the context(s) in which it has a meaning. 135 people participated to this first step and we collected at least $N_i = 68$ annotations per term. In order to make our system robust to the users' bias, we select only the term-context association with a high consensus, i.e., with a high ratio:

$$r(t_i, \psi_k) = \frac{N_i^{(k)}}{N_i} \geq 0.7, \tag{6}$$

where $N_i^{(k)}$ is the amount of annotations for the term t_i in the context ψ_c. Referring to the Eq. 4, we compute the term-context probability by normalizing these ratios, such that:

$$p(\psi_k|t_i) = \frac{r(t_i, \psi_k)}{\sum_{k=1}^{K} r(t_i, \psi_k)}. \tag{7}$$

The selected terms and the relative contexts are listed in Table 1. As assumed, some terms have a semantics in more than one contexts (*calm*, *quiet*, *relaxed*, etc.).

In the second stage of the survey, each pair of terms in the same contexts was proposed to testers, who were asked to annotate their semantic similarity in a given context, with a value ranging from -1 (antonyms) to 1 (synonyms). 170 testers were able to annotate each pair at least 3 times. In order to express the influence of the context in the similarity between terms, we show in Table 2 the similarity we obtained for the terms *calm*, *quiet* and *relaxed* in the two different contexts in which they have a semantics.

3.2. Music Items Semantic Descriptors

In this work we use the public-available MsLite dataset [21], which contains $N = 240$ music excerpts. We created the

Qualifiers' Set $\mathcal{R}$			
Qualifier	Weight	Qualifier	Weight
a little	0.5	moderately	0.6
average	0.7	not	-0.8
completely	1	not at all	-1
considerably	0.9	partly	0.7
extremely	1	quite	0.6
highly	0.9	slightly	0.5
mainly	0.8	very	0.8

Table 3: Qualifiers and mean value weights from [22], scaled to our model

description vectors $\mathbf{d}_1, ..., \mathbf{d}_N$ (Section 2.2) by means of the annotation provided in [4] and we enriched them by means of the procedure explained in Section 2.3 to obtain $\tilde{\mathbf{d}}_1, ..., \tilde{\mathbf{d}}_N$.

3.3. Query Model and Retrieval Model

Our music search system processes text-based queries using the natural language engine proposed in [4]. We selected a set $\mathcal{R}$ of 14 common qualifiers by following the rating scale defined in [22] to allows the user to specifiy the desired intensity of each descriptor in the query. The set of qualifiers and their weights, scaled according to the semantics of our model, are listed in Table 3. In the case no qualifier is specified, we assign the *average* as the default value. We model the query vector $\tilde{\mathbf{q}}$ as explained in Sections 2.2 and 2.3.

Given a query q, and its enriched form $\tilde{\mathbf{q}}$, we implement the musical items retrieval procedure by computing the metric $\xi_{d_i q}$ for each vector $\tilde{\mathbf{d}}_\mathbf{i}$ in the database. In our system, $\xi_{d_i q}$ is defined as the cosine similarity:

$$\xi_{d_i q} = SC(\tilde{\mathbf{d}}_i, \tilde{\mathbf{q}}) = \frac{\tilde{\mathbf{q}}^T \tilde{\mathbf{d}}_i}{\|\tilde{\mathbf{q}}\| \|\tilde{\mathbf{d}}_i\|}. \quad (8)$$

Music items $d_1, ..., d_N$ are ranked according to the score $\xi_{d_i q}$: a higher score express a higher relevance of the item with the respect to the query q.

4. MODEL EVALUATIONS

We compared our system with a LSI-based system [13] and with the system proposed in [4]. The aforementioned LSI exploits co-occurrences of annotations in songs and maps both queries and music description vectors in a h-dimensional reduced space. We used the singular value decomposition to compute a reduced dimensional space with $h = 20$.

The system proposed in [4], on the other hand, is based on a description model that uses two types of descriptors for music: emotional-related descriptors, mapped in the VA plane [8] and non emotional-related descriptors, modeled as dimensional bipolar descriptors. The descriptors and the songs are modeled as normal probability distributions and their similarity is computed as a Bayesian posterior probability.

	DCSM		**LSI**		[4]	
	μ	σ	μ	σ	μ	σ
One context	**6.9**	1.3	6.7	1.4	3.9	2
Different contexts	**6.5**	1.6	6.2	1.2	5.7	1.5
Polysemy	**6.7**	1.5	6.0	1.7	5.3	1.9
All queries	**6.7**	1.5	6.2	1.5	5.1	1.9
Free evaluation	**7.1**	1.2	6.4	1.1	5.0	1.3

Table 4: Mean evaluations for the three semantic models.

The retrieval performances of the three systems have been analyzed with a test that we proposed to 30 subjects. They were asked to rate , in a 9-point Likert scale, the quality of the results for each semantic model in a blind manner: they did not know which model they were been using. The subjects evaluated the retrieved results for nine predefined queries and for a free test of the system.

The mean μ and standard deviation σ of the evaluations are listed in Table 4. The predefined queries have been chosen in order to test our systems with different use cases. In the **one context** use case all the terms in the query belong to the same context. As expected, our approach performances are similar to the LSI, since it is not influenced by the use of contexts. The **different contexts** use case allows the use in the query of terms belonging to different contexts, but without any ambiguities. The **polysemy** use case, instead, also includes ambiguity in the query. As expected, these two uses cases exhibits the advantages in using the DCSM respect to the other approaches. In particular, while evaluations on LSI exhibit a notable drop of performance, the DCSM is proven to be able to effectively disambiguate user queries. We think this is also the reason why the DCSM is preferred by the users during the free evaluation stage.

5. CONCLUSION AND FUTURE WORKS

We proposed a new approach for high-level description of music content, based on contexts able to capture musical aspects and semantic similarity scores between terms in the same context. The semantic relations between descriptors, as well as their contexts membership, have been manually annotated through an online survey. We developed a prototype of a music search engine based on our model and we compared it with the model proposed in [4] and with the LSI model based on co-occurrences [13]. Subjective evaluations show that our model exhibits the best performance, especially for queries containing terms that have a meaning in different contexts.

In future works, we will investigate techniques to automatically compute context membership of the terms and their context-dependent similarities.

6. REFERENCES

[1] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic-description using the cal500 data set," in *Proc. 30th Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR)*, 2007.

[2] P. Knees and G. Widmer, "Searching for music using natural language queries and relevance feedback," in *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics*, pp. 109–121. Springer, 2008.

[3] M. Zanoni, D. Ciminieri, A. Sarti, and S. Tubaro, "Searching for dominant high-level features for music information retrieval," in *Proc. 20th European Signal Processing Conference (EUSIPCO)*, 2012.

[4] M. Buccoli, A. Sarti, M. Zanoni, and S. Tubaro, "A music search engine based on semantic text-based query," in *IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, 2013.

[5] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. 11th International Symposium on Music Information Retrieval (ISMIR)*, 2010.

[6] D. Yang Y. Hu, X. Chen, "Lyric-based song emotion detection with affective lexicon and fuzzy clustering method," in *Proc. 10th International Symposium on Music Information Retrieval (ISMIR)*, 2009.

[7] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011*, 2011.

[8] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[9] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.

[10] L. Barrington, D. O'Malley, D. Turnbull, and G. Lanckriet, "User-centered design of a social game to tag music," in *Proc. ACM SIGKDD Workshop on Human Computation*, 2009.

[11] M. Sordo, F. Gouyon, and L. Sarmento, "A method for obtaining semantic facets of music tags," in *Workshop On Music Recommendation And Discovery (WOMRAD)*, 2010.

[12] C. Laurier, M. Sordo, J. Serr, and P. Herrero, "Music mood representations from social tags," in *Proc. 10th International Symposium on Music Information Retrieval Conference (ISMIR)*, 2009.

[13] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[14] M. Levy and M. Sandler, "A semantic space for music derived from social tags," in *Proc. 8th International Symposium on Music Information Retrieval (ISMIR)*, 2007.

[15] P. Saari, T. Eerola, G. Fazekas, M. Barthet, O. Lartillot, and M. B. Sandler, "The role of audio and tags in music mood prediction: A study using semantic layer projection.," in *Proc. 14th International Symposium on Music Information Retrieval (ISMIR)*, 2013, pp. 201–206.

[16] I. Bartolini, M. Patella, and C. Romani, "Shiatsu: tagging and retrieving videos without worries," *Multimedia tools and applications*, vol. 63, no. 2, pp. 357–385, 2013.

[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[18] R. Miotto and G. Lanckriet, "A generative context model for semantic music annotation and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1096–1108, 2012.

[19] M. M. Bradley and P. J. Lang, "Affective norms for english words (ANEW): Stimuli, instruction manual, and affective ratings," Tech. Rep., Center for Research in Psychophysiology, University of Florida, 1999.

[20] M. Lesaffre, L. De Voogdt, M. Leman, B. De Baets, H. De Meyer, and J. P. Martens, "How potential users of music search and retrieval systems describe the semantic quality of music," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 5, pp. 697–707, 2008.

[21] K. Youngmoo, E. Schmidt, and L. Emelle, "Moodswings: A collaborative game for music mood label collection," in *Proc. 9th International Symposium on Music Information Retrieval (ISMIR)*, 2008.

[22] B. Rohrmann, "Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data," *Project Report. University of Melbourne, Australia*, 2003.