CLUSTERING AUDIO CLIPS BY CONTEXT-FREE DESCRIPTION AND AFFECTIVE RATINGS

Shiva Sundaram, Robert Schleicher, and Julia Seebode
Deutsche Telekom Laboratories,
TU-Berlin, Ernst-Reuter-Platz 7, Berlin, Germany.
{firstname.lastname}@telekom.de

ABSTRACT

In the absence of context, the process of listening to acoustic scenes results in deriving an explicit semantic description and an implicit assessment of its acoustic properties in terms of its affective value. In this work, we mainly exploit the relationship between context-free associations of audio clips containing unconstrained acoustic sources with their affective values for clustering. Using over two hundred clips from the BBC sound effects library, we present a novel, quantitative method to compare the clusters of audio clips obtained using its context-free description with the clusters obtained from their affective measures; namely valence, arousal and dominance. Our results indicate that comparing clusters across representations is a suitable approach to determine an appropriate number of clusters to index audio clips in an unsupervised manner. In this paper we present our findings and examples of the resulting clusters of audio clips.

1. INTRODUCTION

Portable consumer devices like mobile phones nowadays include a recording functionality so that users can easily create their own multimedia content. Blogs and Web 2.0 services like YouTube.com allow for an easy and immediate distribution. Personal web pages with video clips or music for permanent availability and instant accessibility is also commonplace. For both intentions of infotainment and goal-oriented search, an appropriate indexing method is desired, which should ideally work automatically and lead to results aligned with human expectations. Users usually index or label multimedia content with descriptions such as “Bird”, but much of the clips’ attraction arises from its affective quality like “funny”, “touching” and the spontaneous associations one has when watching or listening to them. At the same time, as described in sections 1.1 and 1.3, multimedia content can also be presented to the listener to express something funny or touching in a given context. In this paper, we propose a new way to describe real-life audio material with context-free associations which are not necessarily confined to direct description of the acoustic scene. For example, while “breaking glass” (CLEAR BROKEN GLASS from the BBC Library) can be understood as the direct description of the clip, the context free association can be “accident” or “mishap”, which can then be used for clustering, indexing and searching. We mainly address unsupervised clustering based on affective quality and the spontaneous associations the listeners reported. We present a novel approach to select a suitable number of clusters by comparing clustering results across two different representation spaces.

This paper is structured as follows: first, we will address why affective quality is relevant for multimedia clustering, searching and indexing. Next in section 1.2 we will give a brief overview on various human descriptions for audio and third, we will point out that audio material pre-processed in this way can also serve as a template for pre-selecting appropriate samples for the design of auditory icons. In section 2 of this paper, we describe our own experiment where we asked subjects to rate selected clips from the BBC sound effects library on affective dimensions and refer their spontaneous association with it. In section 3 we report some results which show two what extent clustering according to both types of descriptors (affective and semantic) yields to comparable results. Finally, in section 4 we present conclusions and further discussions on the work presented here.

1.1 Affective Quality

Next to acquisition of information, entertainment is presumably one of the main reasons people access multimedia content. *Entertainment* in this context is not confined to amusement or pleasure, but also includes the deliberate exposure to material that is scary or irritating to a certain extent. From a psychologist’s point view such behaviour can be interpreted as a conscious attempt to modulate one’s own mood or *emotion regulation* [3]. With this purpose in mind, it becomes apparent that automatic multimedia search should also aim at indexing their content regarding to its affective quality. A point that has been of interest in the research community for a while [12].

One very general approach to quantify and compare the affective impact of a stimulus is its value on the three dimensions of valence that is sometimes also called *pleasantness* ranging from *unpleasant* to *pleasant*, arousal from *calm* to *aroused* and dominance from *being in control* to *being controlled*. Quantification according to these dimensions has been applied to visual as well as acoustic material by Bradley and Lang [6, 11] which use a pictogram-based scale, the *Self-Assessment-Manikin* SAM,[5] to let users rate the perceived affect. The SAM is briefly described in the method section of this article.

These dimensions are very general and can be applied to all kind of stimulus material. At the same time, the advantage of generalizabilty also brings along the disadvantage of a certain lack of specifity, implying that stimuli which cause different emotions like anger or fear will get quite similar ratings, namely negative valence and high arousal [13]. One attempt to overcome this drawback is to map emotion categories onto the SAM values [16]. For clips with a distinct semantic content this may be feasible, but for potentially ambiguous material like real world audio clips, a distinct classification may be less successful. If the value on the universal dimensions of valence, arousal and dominance can be seen as one very general approach to quantify an audio clip, the individual associations each user has while listening to it

 ISSN 2076-1465

can be regarded as the other extreme, as they are determined to great extent by this person's individual interpretation and memories. Emotional impact not only triggers attention, but is also a strong memory enhancer, thus making it likely that material with a strong emotional connotation will also be rich in associations on the user's side. Both properties, attention-grabbing and evoking of clear associations are also named as desired features of auditory icons which is why we chose this as an additional exemplary application.

1.2 Semantic Descriptions

Due to the inherently rich content of audio and multimedia, there are a variety of language-level descriptions for it. A common example is the exact description of the content (e.g., `CLEAR BROKEN GLASS`). There have been many successful approaches to using this form of description [2, 7, 14]. These examples also exploit category-based taxonomy for indexing. Alternatively, onomatopoeia-based descriptions has also been looked at [17].

In this work, we are mainly interested in spontaneous associations to audio clips (using descriptive words such as "funny" or "touching") that are typically available in the form of reviews or users' comments in Web 2.0 applications. This form of specific description does not necessarily describe the content exactly, but describes the user's association after listening to the acoustic sources in the clip. Also, the experience of listening to the clip is measured in a generalized manner using the affective values. We believe that the two forms of subject-based inputs help cope with each other's limitations of generality and specificity. We use the associations to derive a semantic representation of the audio clip using latent semantic indexing described later in section 2.2.

1.3 Auditory Icons

Bill Gaver [10] established the concept of auditory icons as caricatures of sounds that appear in the real world. He emphasized their most important advantage that people listen to those sounds in their everyday life and thus are used to this kind of auditory information. In addition the mapping between a certain sound and the event or object it represents is not arbitrary as it would be with the use of artificial sounds. Therefore the sounds can intuitively be mapped as analogies to the actions or events. For instance, deleting a file can be mapped to the sound of a rumpled piece of paper being thrown into the recycling bin. Finding suitable sounds that are associated to certain events and objects in human-computer interfaces is a challenging task for the design of auditory icons as not all events in this context produce a sound that is directly related. Metaphorical mappings without ambiguity need to be found in these cases. The stronger existing associations are the better are learning and retention rates of sound event pairings [15]. Hence, auditory icons can be a powerful approach to provide information about an event or object in a human-computer interface or in a context set by the content. Given that their acoustic meaning evokes clear and distinct associations in the user. The challenge is to find clips that have this property.

2. METHOD

To obtain standardized data set that comprises both affective ratings and context-free association we have undertaken our own controlled data collection procedure described below.

2.1 Data Collection

For collecting context-free associations and the corresponding affective ratings we used a selection of 219 audio clips from the BBC Sound Effects Library [1]. The audio clips were originally designed to mix audio tracks and consist of natural, unconstrained audio recorded in real environments. To obtain reliable results, we wanted to have an experimental time to rate all the sounds in less than one hour. We split our selection of clips and conducted two similar experiments with about 110 clips each. In addition to the BBC clips we included 5 clips from the IADS database [6] to compare the ratings of our subjects to the values obtained in the IADS database in each test. This was used to cross check affective ratings of our samples with the reference samples in [6].

Altogether, 64 students from the local university campus (35 male, mean age 26.5, std. 4.2) participated in the two experiments. Upon arrival, subjects were asked to fill in a general demographic questionnaire. Next they were seated in an anechoic chamber and introduced to the general set-up and procedure, including the computerized version of SAM. In this questionnaire, valence also sometimes called *pleasantness* is depicted with a face varying from happily smiling to unhappily frowning in the SAM rating scale. A typical stimulus would be baby laughter compared to crying. Arousal is represented by a pictogram that ranges from an excited figure with eyes wide open to a relaxed and sleepy character. Dominance is illustrated as a manikin of increasing size that for high dominance values is almost covering the complete frame. The bipolar nature of all dimensions also allows classification of stimuli that maybe neutral with regard to one aspect. We used the original instruction [5] translated to German where it is explicitly stated to mark the medium value of 5 on the 9 step scale in case of indecision. To let the subjects practise the rating and type their associations (which could only be done once the sound was played completely) an exemplary sound file was used. While listening to this example clip, subjects were said to adjust the loudness to their individual preference. Once a preferred level was set, it was recommended not to change it during the experiment. As a last step before the actual experiment started, the lights were dimmed to avoid visual distractions and let the subjects focus on the sound stimuli. Sounds were presented using a Sennheiser 280HDPRO headphone and a Fujitsu Siemens S7110 Laptop with an external Millenium HP3 headphone amplifier. After listening to all the sound clips in randomized order, subjects were asked about any remarks or observations which were noted by the experimenter. Finally, they were paid a sum of 15 Euros for their effort and thanked. Altogether a session lasted about 60 - 90 minutes.

At the end of the data collection procedure each of the 219 audio clips contains on average of 16 context-free text associations which underwent a standard manual text normalization procedure. As a first step spelling mistakes were corrected, special characters (like dots, commas etc.) were removed and all letters were transformed to lower case. Additionally, all function words (like articles, prepositions etc..) were removed as they don't convey any lexical information. Afterwards, we normalized all verbs to their infinitive, all nouns to their singular and all adjectives and adverbs to their word stem. As a final step we replaced synonymous words, e.g. the words "bothersome, annoying, bother" were all con-

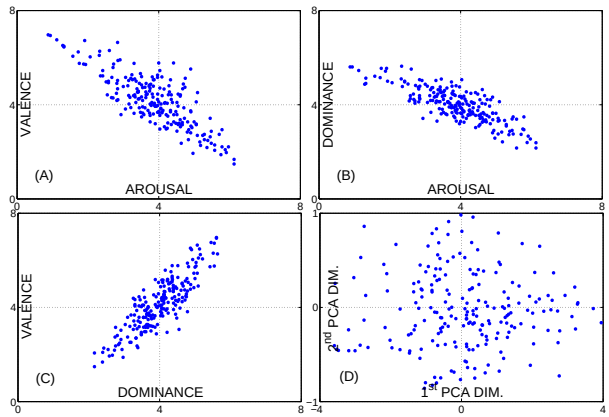

Figure 1: (A-D) Scatter plots of audio clips by affective dimensions. (D) the 1st and 2nd dimensions after PCA. Note: Each audio clip is a single point.

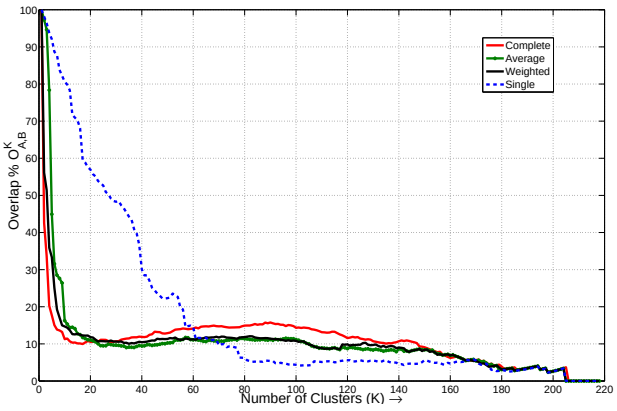

Figure 2: Across-space overlap of clusters as a function of number of clusters (K) for different clustering criteria.

sidered "annoying" because of the most frequently named word (in this case "annoying"). In addition to the normalized text association, for each clip, a 3-tuple comprising the affective ratings was also obtained.

2.2 Clip Representation

Using the subject-based data, we derived two forms of representations for each audio clip: a vector in the latent space using the text associations and a three dimensional representation after principal component analysis (PCA) of the 3-tuple affective measures. The representation in the latent space was obtained by latent semantic indexing (LSI) (see [8]). Typically, LSI of text documents begins with term-document frequency counts between the document (the text or the associations of a clip) and the terms (a selection of meaningful words). Many state-of-the-art text document indexing and retrieval approaches fundamentally rely on this measure. In addition to direct term-document frequency counts, a entropy-based measure is also used to weigh the entries in the term-document frequency matrix [4]. In LSI, the final representation is derived by dimension reduction using singular value decomposition (SVD). From the complete collection of associations, a total of 1634 unique words were hand selected after a typical text normalization procedure. The associations corresponding to each clip were used as text document and the hand selected words as terms. The approach has many advantages in the application of semantic retrieval of text documents, these can be found in [4]. This approach is suitable for the present investigation because it is completely non-probabilistic and makes no model-based assumptions about the underlying data. It is also applicable to small to medium size data.

The scatter plot of the audio clips in terms of the valence-arousal, arousal-dominance and valence-dominance is shown in figure 1. It can be seen that scatter of points form an elliptical shape indicating a correlation between the dimensions also normally observed in [6, 11]. Therefore, to orthogonalize the dimensions and retaining only the required information, we performed a PCA and reduced the dimensions to two. This is illustrated in the lower-right sub-figure (D) of figure 1.

We are able to derive vectors for audio clips in two different representation spaces: using the text-based context free associations and using the SAM ratings. The relationship between the two representations can be better understood by unsupervised clustering. This is described in the next section.

2.3 Clustering in Latent and PCA space

Clustering allows us to impose a meaningful structure to the audio clips in a given representation space. In this work, due to the processes of human sound identification, an implicit relationship (or mapping) between the affect of an audio clip and its context-free association exists. This relationship can be direct but surjective because a point in the latent space is related to a point in the PCA space as they both represent the same audio clip. Notwithstanding the fact that two distinct audio clips may have exactly the same association or they may have exactly the same affective values, similar audio clips shall lie in the vicinity of each other in the representation spaces. Therefore imposing a structure by clustering is of interest to group and organize similar clips together. Additionally, within the framework of multimedia content retrieval, an unknown audio clip can be indexed either using its affective score or by descriptions and subsequently similar clips from the cluster it belongs to can be retrieved for a user. In this work, the listening experience is quantitatively captured by its affective ratings and mapped to the PCA space and the context free association is captured by the text description and later represented as a vector in the latent space.

When considering clustering two factors need to be considered: the number of clusters and the clustering method. Ideally, due to the relationship between the affective values of audio clips and the association, it is desirable that on an average, the resulting clusters in one space overlap with the clusters in the other space. For reasons mentioned in section 3, in this work, we use hierarchical clustering in which the resulting dendrogram needs to be cut at an appropriate level for selecting the number of required clusters. For this purpose, we propose to use across-space overlap between clusters. A follow up question on this that needs to be addressed is what is a suitable metric for measuring the overlap between the clusters formed in different representation spaces. This is described next.

Let $s_1, s_2, s_3, \ldots, s_N$ be N audio clips. Depending on the application in question, these clips can be represented as single vectors in multiple spaces. In this work, we are interested in two: one derived from affective values and the other is the latent space derived from associations (denoted as $\mathscr{A}$ and $\mathscr{B}$). These clips can be clustered in an unsu-

Cluster V=5.63, A=2.45	GREAT-TIT-TEACHER-SONG-B2, ROBIN-SONG-BB RANGANATHITTU-BIRD-SANC2-B2, BIRD-WILLOW-WARBLER-BB, BRITISH-CHAFFINCH-BB, BRITISH-HOUSEMARTIN-BB, BRITISH-BLACKCAP-BB, CHIFFCHAFF-BIRD-B2, ROSEATE-COCKATOOS-AUSTRL-BB, COCKEREL-AND-HENS-GRASS-BB, PEREGRINE-FALCON-CALLS-B2, IBERIAN-MARSH-FROG-BB GARGANEY-BB, GOLDCREST-OTHER-BIRDS-B2, 1-DIESEL-LOCOMOTIVE-PASS-B2, MALLARD-MALE-FEMALE-B2, MISTLE-THRUSH-SONG-B2, MOORHEN-BB, CITY-PARK-DUCKS-POND-BB, CAMEROON-RAINFOREST-BB, ADULT-ROOK-CALLING-B2, RUDDY-SHELDUCK-S-ASIA-BB, 2-ADULT-SPARROWS-SING-B2, BLUSTERY-WIND-BEACH-BB.
Cluster V=3.49, A=4.23	ANTARCTIC-WHITE-OUT-WIND-BB, CASTLE-RAIN-THUNDER-BB, REAR-DEBRIS-FALLING-BB, GALE-HEARD-FROM-INSIDE-BB, HAIL-ON-UMBRELLA-BB, POLAR-WIND-BB, HEAVY-RAIN-ON-CAR-INT-BB, RAIN-ON-CORRUGATED-IRON-BB, RAIN-ON-WINDOW-WATER-BB, RAIN-AND-DISTANT-THUNDER-BH, APPROACHING-THUNDER-BH.
Cluster V=3.62, A=4.27	CAR-WINDOW-DWN-UP-SQUEAK-BB, PARIS-TRAIN-PLATFORM-B2, COMMUTER-RAILCAR-RUN-B2, RAILWAY-STATION-BB, LONDON-SUBWAY-ARRIVES-01-B2, DIESEL-TRAIN-PASSES-INT-B2, BUDAPEST-TRAM-STATION-B2.
Cluster V=3.06, A=4.67	CLEAR-BROKEN-GLASS-BB, CROCKERY-BROKEN-BB, GHOSTLY-FOOTSTEPS-CHAINS-BH, LARGE-GLASS-CRASHES-BB, MED-TO-LRG-GLASS-BB, NAIL-PULLED-OUT-01-B2.
Cluster V=3.36, A=4.34	ALARM-TICKING-BB, CHIMING-CLOCK-1-OCLOCK-BB DIAL-PAY-PHONE-PIPS-BB, ROTARY-PAY-PHONE-DIAL-BB PAY-PHONE-CALLBOX-RING-1-BB, PHONE-CALLED-PARTY-RINGS-BB, PHONE-PABX-RINGING-TONES-BB, PHONE-STD-RINGS-ANSWER-2-BB, PHONE-STD-RINGS-ANSWER-1-BB, BT-TELEPHONE-ENGAGED-BB.
Cluster V=4.79, A=3.55	BUSY-RESTAURANT-CHATTER-B2, MICROWAVE-LOAD-RUN-BB, PARIS-CAFE-INTERIOR-B2, PUBLIC-BAR-QUIET-CROWD-BB, QUIET-SMALL-RESTAURANT-B2.

Table 1: Typical examples of clusters and their average affetive values (valence "V" and arousal "A") obtained by hierarchical clustering of audio clips for $K = 60$.

pervised manner using a similarity metric. Let this result in $\{C_1^A, C_2^A, C_3^A, \ldots, C_{K_A}^A\}$ and $\{C_1^B, C_2^B, C_3^B, \ldots, C_{K_B}^B\}$ mutually exclusive clusters. Without loss of generality, as we are clustering the same set of audio clips, it is convenient to assume K_A=$K_B = K$. In addition to choosing a suitable clustering method, quantitatively defining correct clustering between elements is not trivial. Typically, if labelled data (as preformed categories or clusters) were available, it would be possible to compare the results of unsupervised clustering and subsequently define correct clustering. In this work we are only focussing on using the context-free associations and affective ratings of audio clips.

Correct clustering across representation spaces between elements can be defined in the following way: if two or more elements are simultaneously grouped together in a cluster in space $\mathscr{A}$ and also grouped in a cluster in space $\mathscr{B}$, then they have been correctly clustered. That is, in case of two elements, if for any j and k if s_j and s_k simultaneously belong to p^{th} cluster in $\mathscr{A}$ and a q^{th} cluster in $\mathscr{B}$, then they have been correctly clustered. Even for two selected elements, this procedure is tedious for a reasonable number of clusters. Determining correct clustering also applies to permutations of more than two elements and therefore a direct procedure that goes through all combinations of elements and clusters can be computationally prohibitive. To alleviate this issue, we propose to use a matrix approach.

Let $s_1, s_2, s_3, \ldots, s_N$ be grouped into K clusters in space $\mathscr{A}$. Let A^K be a $N \times N$ triangular matrix with entries a_{ij}^K, where

$$a_{ij}^K = I_q(s_i) \cdot I_q(s_j) \quad \forall\, 1 \leq q \leq K \,\forall\, i < j,\, 1 < j \leq N \tag{1}$$

In the above equation, $I_q(\cdot)$ is an indicator function. Matrix A^K is obtained by applying a fixed clustering procedure to he vectors in space $\mathscr{A}$. In our work, a matrix B^K is also obtained by using equation for space $\mathscr{B}$. Subsequently, for K clusters, the across-space *overlap* between the resulting clusters in $\mathscr{A}$ and $\mathscr{B}$ can be defined as:

$$O_{A,B}^K = \frac{A^K \odot B^K}{A^K \oplus B^K} \quad \text{Where } 1 \leq K \leq N \tag{2}$$

Here,

$$A^K \odot B^K = \sum_{i,j}^{N} a_{ij} \cap b_{ij} \tag{3}$$

is the intersection of matrices A^K and B^K and,

$$A^K \oplus B^K = \sum_{i,j}^{N} a_{ij} \cup b_{ij} \tag{4}$$

$\forall\, i < j,\, 1 < j \leq N$, is the union of matrices A^K and B^K. It is straightforward to see that for a given clustering procedure, the overlap $O_{A,B}^K$ can be determined for different number of clusters K and this measure can be used to determine *correct* clustering across-spaces.

3. RESULTS

In the information retrieval community a variety of clustering methods for data have been proposed and studied [9]. For this work, we use hierarchical agglomerative clustering because they do not impose a probabilistic model on data and are practical for experiments on limited data set. Figure 2 illustrates the overlap percentage (equation 2) for four different merging criteria. These were selected because for the current application they resulted in monotonic clusters. In the figure it can be seen that for small number up to about 60 clusters, the area under the curve for single-linkage criterion is larger than the area for the other criteria. This indicates that in this region it performs best. However from about 60 and above, the area under the curve for complete-linkage is the largest. This observation is consistent in comparison with other clustering criteria also. While the exact value of number of clusters is data dependent, it is easy to see that this method can be unequivocally applied to larger datasets. A large gap between two successive combinations in agglomerative clustering is a good point to cut the dendrogram to obtain a partition of mutually exclusive or disjoint clusters. As indicated in figure 2, $K = 60$ is a reasonable choice. The resulting clusters also indicate that at this value, the number of elements under each cluster is less skewed. Examples of the resulting clusters are indicated in Table 1. The table also indicates the average valence (V) and arousal (A) values for the cluster. A closer look at the relative values is interesting. The first cluster of bird sounds is pleasant but not very arousing. In comparison to this, the fourth cluster in the table with breaking sounds is less pleasant but more arousing.

4. CONCLUSION AND DISCUSSION

In this work, we have presented our results on unsupervised clustering of audio clips in two representation spaces: affective values after PCA, and latent space from context-free associations. A controlled data collection procedure was undertaken to obtain reliable subject-based data for affective values and context-free associations for a selection of clips from the BBC Sound Effects Library. A novel approach to quantitatively measure the overlap between clusters across spaces was also presented. This measure was used to select an appropriate clustering procedure and the number of clusters. The results show that the affective rating scheme based on the dimensions of valence, arousal and dominance can be deployed for this purpose, especially for unconstrained audio content that is typical in Web 2.0 applications. Comparing clusters across-spaces, is a viable approach to select an appropriate number of clusters for unsupervised clustering.

Multimedia retrieval using affective quality has gained interest from other researchers [18], however we believe that our approach provides refinements to the current state of research: the Self-Assessment Manikin (SAM) [5] represents a convenient and widespread measurement tool that is well established in emotion research, thus linking our results to findings in that domain. Additionally, there is standardized stimulus material available, auditive (the aforementioned IADS) as well as visual (International Affective Picture System, by the same authors) enabling comparisons across modalities. While the affective quality might trigger the initial reaction to a stimulus in terms of approach/interest or withdrawal/avoidance, the mere affective description might be too unspecific in some cases as described in section 1.1, which is why we propose to enrich them with the individual context-free associations in contrast to existing frameworks [18]. It is certainly the case that affective dimension values alone are not sufficient to describe/index multimedia material, but the descriptors are usually implicit in the user tags. If we can then augment it with our estimated affective value, we can utilize the combined information for better content retrieval. For example, the relatively low valence value of the 4^{th} cluster obtained in 1 shows the sound of a "breaking glass" is indicating something unpleasant. This together with the spontaneous association "mishap" makes it an appropriate candidate to indicate something the crashing of an application as an auditory icon. As a part of our ongoing work, we will continue to obtain subjective measures for all the clips in the BBC library which has emerged as a standard for training and testing new audio indexing systems [2, 14]. Not only will this help us and the research community compare various retrieval mechanism, it will also help us identify acoustic signal correlates that can help predict/estimate affective values for an unknown clip.

REFERENCES

[1] The BBC Sound Effects Library - Original Series. *http://www.sound-ideas.com*.

[2] L. Barrington, A. Chan, and D. T. land G. Lanckriet. Audio Information Retrieval using Semantic Similarity. *Proc. of ICASSP, Honolulu, Hawaii*, 2007.

[3] J. S. Beer and M. V. Lombardo. Insights into emotion regulation from neuropsychology. In J. J. Gross, editor, *Handbook of emotion regulation*, pages 69–86. Guilford Press, New York, NY, 2007.

[4] J. Bellegarda. *Latent Semantic Mapping: Principles and Applications*. Morgan and Claypool, 2007.

[5] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, Mar 1994.

[6] M. M. Bradley and P. J. Lang. The international affective digitized sounds (2nd edition; iads-2): Affective ratings of sounds and instruction manual. *2nd Ed. Affective ratings of sounds and instruction manual (Technical report B-3). University of Florida.*, 2007.

[7] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard, P. Herrera, and N. Wack. Nearest-Neighbor Generic Sound Classification with a WordNet-based Taxonomy. *In Proc. 116th Audio Engineering Society (AES) Convention, Berlin, Germany.*, 2004.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. Landaurer, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 6(41):391–407, 1990.

[9] R. Duda, P. Hart, and D. Stork. *Pattern Classification*, volume 2nd edition. Wiley-Interscience, 2000.

[10] W. Gaver. Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction*, 2(2):167 – 177, 1986.

[11] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (iaps): Affective ratings of pictures and instruction manual. technical report a-6. *University of Florida, Gainesville, FL*, 2005.

[12] R. W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, 1997.

[13] R. Schleicher and N. Galley. Continuous rating of experienced affect while watching emotional film clips. *Psychophysiology*, 46(suppl.1):s51, 2009.

[14] M. Slaney. Semantic-Audio Retrieval. *International Conf. on Acoustic Speech and Signal Proc. (ICASSP), Orlando, USA.*, pages 13–17, May 2002.

[15] K. Stephan, S. E. Smith, R. L. Martin, S. P. Parker, and K. I. McAnally. Learning and Retention of Associations between Auditory Icons and Denotative Referents: Implications for the Design of Auditory Warnings. *Human Factors*, 48(2):288–99, 2006.

[16] R. A. Stevenson and T. W. James. Affective auditory stimuli: characterization of the international affective digitized sounds (iads) by discrete emotional categories. *Behavior Research Methods*, 40(1):315–21, Feb 2008.

[17] S. Sundaram and S. Narayanan. Classification of sound clips by two schemes: Using Onomatopoeia and Semantic labels. *Proc. of ICME, Hannover, Germany*, June 2008.

[18] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457, Feb 2008.