Journal of the Audio Engineering Society
Vol. 64, No. 7/8, July/August 2016 (© 2016)
DOI: http://dx.doi.org/10.17743/jaes.2016.0024

On the Impact of the Semantic Content of Sound Events in Emotion Elicitation

KONSTANTINOS DROSSOS,[1] *AES Member* , **MAXIMOS KALIAKATSOS-PAPAKOSTAS**[2],
(konstantinos.drossos@tut.fi) (maxk@mus.auth.gr)

ANDREAS FLOROS,[3] *AES Member*, **AND TUOMAS VIRTANEN**[1]
(floros@ionio.gr) (tuomas.virtanen@tut.fi)

[1]*Audio Research Group, Dept. of Signal Processing, Tampere University of Technology, Tampere, Finland*
[2]*Cognitive and Computational Musicology Group, Dept. of Musical Studies, Aristoteles University of Thessaloniki, Thessaloniki, Greece*
[3]*Lab. of Audiovisual Signal Processing, Dept. of Audiovisual Arts, Ionian University, Corfu, Greece*

Sound events are proven to have an impact on the emotions of the listener. Recent works on the field of emotion recognition from sound events show, on one hand, the possibility of automatic emotional information retrieval from sound events and, on the other hand, the need for deeper understanding of the significance of the sound events' semantic content on listener's affective state. In this work we present a first, to the best of authors' knowledge, investigation of the relation between the semantic similarity of the sound events and the elicited emotion. For that cause we use two emotionally annotated sound datasets and the Wu-Palmer semantic similarity measure according to WordNet. Results indicate that the semantic content seems to have a limited role in the conformation of the listener's affective states. On the contrary, when the semantic content is matched to specific areas in the Arousal-Valence space or also the source's spatial position is taken into account, it is exhibited that the importance of the semantic content effect is higher, especially for the cases with medium to low valence and medium to high arousal or when the sound source is at the lateral positions of the listener's head, respectively.

0 INTRODUCTION

Hearing and vision are the two mostly employed sensory modalities for communication [1, 2]. Through the corresponding communication channels, i.e., audio and visual, we communicate with other people, express our thoughts and ideas, entertain ourself and other persons, and perceive knowledge for our surroundings and environs. Along with these we also discern, transmit, and elicit emotions [3–5]. Focusing on sound, it is reported that there are three types of audio stimuli: (i) speech, (ii) music, and (iii) non-verbal and non-musical sounds termed as general sounds, everyday sounds or sound events (e.g., environmental sound events such as a car passing by, dog barking, etc.) [6–9]. For the former two there is a well structured research background and the corresponding scientific fields have to demonstrate high achievements. For example, there are various published works regarding speech and music related processing and recognition tasks [10–12]. There is also an increased interest on the emotion recognition from speech and music. Various published works investigate the relation of the aforementioned two types of audio stimuli with the elicited motion on the listener [3, 13, 14]. But, speech and music are the smaller portion of the total heard audio stimuli. The bigger portion is occupied by general sounds, i.e., sound events [4].

Emotion recognition from sound events is a recent field of study with few produced results [15]. The existing published works are focusing on the investigation for a systematic relation between the acoustic cues of the audio stimuli and the conveyed emotion to the listener. However, the technical characteristics (both signal/stimulus and source related characteristics) are just one of the factors that can affect the listener's emotional state [1]. Specifically, the listener's individuality (i.e., his personality, background, culture, past life, and others) and the semantic content of the general sound can have an impact of the elicited emotion.

With respect to existing datasets, e.g., International Affective Digitized Sounds (IADS) where each sound has been rated by at least 100 people, the impact of the individuality of each annotator is greatly decreased since the annotations are average across all annotators. The remaining factors affecting elicited emotions are the technical characteristics and the overall semantic content. The latter can

be used to compute *causal* and *semantic* similarity, as has been differentiated by existing works focusing on the similarity of general sounds [16–18]. Briefly, causal similarity refers to the actions that produce a sound, indicated by its describing verbs, whereas semantic similarity refers to the sound sources, indicated by the nouns describing the sound.

The impact of the semantic content of sound events to the listener's emotion has not been studied. In particular, questions like: "What is the importance of the semantic content of a sound event with respect to the elicited emotion?"; "Do sound events with similar semantic contents elicit similar emotions?"; "Is it the knowledge that the clang is a gun, or is it the technical characteristics of audio stimulus that more strongly affect the listener?"; or "Does the semantic content of a sound event have a more profound impact on the elicited emotion than its actual acoustic cues?" are yet to be answered.

The present work focuses on the impact of the semantic content of sound events to the listener's elicited emotion. According to the above research questions, we present a first approach towards investigating whether two sound segments emerging from sources named with semantically similar names, can elicit a similar emotion. More specifically, the work at hand examines the semantic similarity of sound events that produce similar emotional states to the listener. We utilized two datasets with emotionally annotated sound events. One without spatial information of the source, i.e., the IADS [19], and the Binaural Emotionally Annotated Digital Sounds (BEADS), which consists of binaural rendered (i.e., with spatial information) versions of sound events present in IADS [20]. Both of these datasets employ the widely adopted Arousal-Valence (AV) space with clustering according to Self Assessment Manikin (SAM) values [21]. The semantic similarity was measured by using the well established Wu-Palmer similarity measure [22]. The rest of the paper is organized as follows: in Section 1 we present a brief overview of the related works focused on emotion modeling and annotation, emotion recognition from sound events, and semantic similarity measurement based on WordNet [23]. Section 2 presents the experimental procedure followed by the illustration of the obtained results and their discussion in Section 3. Section 4 concludes the paper and proposes future enhancements for the current field of research.

1 RELATED WORK

What emotions exactly are is a question that is still debated between experts in the relevant fields. Nevertheless, emotions can be modeled by using two approaches: (i) discrete, and (ii) continuous models [1]. The former category includes models that use discrete verbal descriptions in order to model emotions. The most typical representative is the well known basic emotions model [24]. The latter category includes models that approach emotion as the resultant of N discrete affective dimensions with typically $N = 2$, i.e., Arousal-Valence. The latter category seems to be preferred in engineering related works as it provides a reduced ambiguity on the annotated emotion. With the discrete emotions models there is a reported problem related to the perceived or intended meaning of the employed word (e.g., "Happiness" versus "Happy"), and the ability for later mapping of values to clusters representing or assigned to specific verbal descriptions of emotions [1, 14]. For emotional annotations using a continuous model the SAM [21] has been developed, with which a person can quantitatively annotate his/her affective state. SAM consists of a series of drawn manikins representing different values of corresponding affective states and a set of intermediate values, i.e., in between the figures. Altogether there are nine available choices for emotional annotation with one representing the lowest value and nine the highest.

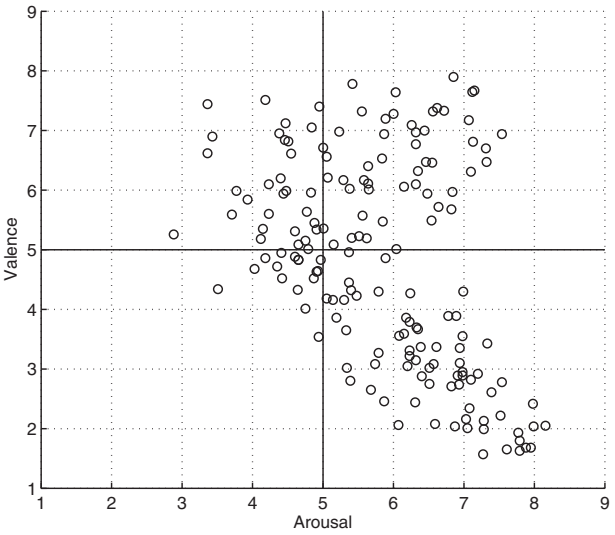

Fig. 1. Scatter plot of the IADS AV annotations

Another important aspect in the research centered on emotional information retrieval from sound events is the available datasets. Currently there are three freely emotionally annotated datasets with sound events. In a chronological order, the first one is IADS [19]. It consists of 167 sound events, emotionally annotated for arousal, valence, and dominance and with content annotation, i.e., for all sound events in the IADS dataset there is a string representation of the content, e.g., "dog," "enginebreak," "busysignal," etc. In Fig. 1 the scatter plot of the IADS annotations in the AV space is illustrated. The second is the Emotional Sound Database (ESD) [5]. It consists of 360 sound events, emotionally annotated for arousal and valence and also with content annotation. The third is the BEADS dataset [20]. It consists of 32 sound events, binaural rendered for 5 angles (i.e., $32 \times 5 = 160$ in total sound events), emotionally annotated for every angle, and also with content description. BEADS is based on IADS and is the only existing emotionally annotated sound events dataset containing audio stimuli with spatial information.

Even though there are not many published works focusing on sound events emotion recognition [15], most of them use AV and/or SAM and at least one of the available datasets with emotionally annotated sound events. The AV model is employed in [1, 15, 5, 25] and the SAM model is employed in [19, 20, 4]. ESD is used in [15] and [5] and

most of the other works employ IADS and/or BEADS. The semantic content seems not to be tackled in many of the existing works on the emotional information retrieval from sound events, even if some of them mention its importance in the resulting emotion recognition. For example, in [1] the authors have presented a new approach to the acoustic ecology by expanding it, i.e., affective acoustic ecology, and re-defining the sound event in the scope of the emotionally enhanced acoustic ecology. This new definition consists of the semantic content along with the source's spatial position, the waveform, and the duration of the sound event. In addition, in the same work it is stated that the semantic content of the sound events might affect the arousal of the listener, as indicated by the presented results. In [4] the authors investigated the impact of sound source angular position to the listener's affective state. They discussed the obtained affective state ratings according to spatial location of the source and the expansion of the research on the emotion recognition from sound events by also taking into account the semantic content of the audio stimuli. Finally, in [15] is presented a study on the common characteristics of music, speech, and sound events and how these can affect the emotions of the listener. The results indicated that a cross-domain (i.e., for speech and/or music and/or sound events) arousal and valence estimation is feasible, but it is also hard, as the authors say, in terms of obtaining a standard feature set that could achieve equally well in a cross-domain scenario. Furthermore, the authors seem to strengthen the need for semantic analysis of sound events by bringing forward the fact that different kinds of general, or natural, sounds were employed in speech and music for expressive functions.

In order to measure the impact of the semantic content a framework that semantically connects and/or represents the notions/senses of words must be employed. One such framework that is also widely adopted is WordNet [23]. It was developed in order to provide a combination of traditional lexicographic information with a computer interface for automated or programmable access and process of stored information. WordNet organizes words appearing in the English language in a hierarchy (tree-like) structure where the higher nodes in the structure represent more abstract or general meanings/senses and deeper nodes (or leaves) are more specific meanings/senses [26]. In addition, WordNet contains information for what part of speech each word is, which is not necessarily unique. For example, the word "back" can appear as noun or as verb [23]. The ambiguity of the word's meaning is resolved by a proper representation of the word. Thus, and according to the previous example, the word "back" has nine meanings as a noun, ten as a verb, three as an adjective, and six as an adverb[1]. In the application programming interface (API) of WordNet, the selection of a specific meaning is implemented by the following representation: $< word > . < part\_of\_speech > . < order >$. So, if the second meaning of the word "back" as verb is needed the correct representation in the API is "back.v.02." The precise word meaning obtained by the aforementioned process will be called synset for the rest of the paper, following the WordNet terminology.

On top of the WordNet structure, methods have been developed in order to measure the semantic similarity. These can abstractly be grouped in two classes: one that uses path-based and another that uses information content-based measures [27]. The former class of methods includes those that use the distance and/or the path between two synsets in the WordNet tree-like structure while on the latter are those that use the information content (IC) for each synset. The IC-based methods are heavily relying on the corpora and the calculation of the IC [26, 27], a fact that makes it possible to obtain different similarity measure with different IC corpora. For that reason, in the present study, we focused on the methods that use path-based calculations.

Path-based similarity measures for WordNet are: (i) *Shortest path similarity*, which takes into consideration only the shortest path between two synsets and has a range of values [0, 1], where 0 indicates no connection between the two synsets and 1 the same synset/meaning; (ii) *Wu-Palmer's similarity*, which is a weighted distance measure that takes into account the positions of the synsets and the position of their most specific (i.e., deeper) common ancestor in the hierarchy and has a range of values (0, 1], with 1 indicating the same synset, and the lower the value the less similar the synsets are; and (iii) *Leacock-Chodorow similarity*, which is similar to (ii) but takes into consideration the depth of the taxonomy into which the synsets are found and has a range of values [0, 3.7).

We use the Wu-Palmer semantic similarity measure due to: (i) having an already weighted value, and (ii) allowing the calculation of semantic similarity of nouns and verbs, according to the current implementation of the nltk python package and Wu-Palmer function [28]. As data we employed the synsets that correspond to the names of sound events that are contained in the IADS and BEADS datasets by manual selection of proper order for the synsets in WordNet. We used both datasets in order to investigate not only the semantic similarity with different affective states but also any underlying relation of the semantics and source's spatial position. As different affective states we considered all pairs from SAM's manikin choices and for the dimensions of AV (that are common in both datasets).

2 FOLLOWED PROCEDURE

The followed procedure had three phases. The first one regarded the creation of the synsets that would be utilized in the semantic similarity measurement. The second considered the clustering of synsets according to emotion annotation values, and the third was the actual semantic similarity measurement. These phases are presented thoroughly in this section.

[1] accessible at: http://wordnetweb.princeton.edu/perl/webwn?s=back&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=.

2.1 Creation of Synsets

As mentioned in Sec. 1, a synset consists of the word, its part of speech, and the proper order in the WordNet structure. Both employed datasets provide a list of the content for each sound event, i.e., a verbal description of the content in each sound event. Based on that list, we manually selected the proper order in the WordNet though its online search engine[2]. Most of the words in the content descriptions were nouns except the word "busy," for the sound event from IADS with content annotation of "busysignal." Some content annotations in both datasets were single words, e.g., "dog," "cat," and some double words, e.g., "engine break" written as "enginebreak." The single words were transformed to single synsets and the double words to a set of two synsets. Therefore, if a word in the content description was apparent in the WordNet, then a single synset was employed, while otherwise, two synsets were used. After the creation of synsets, each sound event in each dataset was described by its emotion annotation values, i.e., AV values provided by the datasets, and a synset (if its content description was a single word) or a set of two synsets (if its content description was a dual word). A full list of employed synsets can be provided upon request.

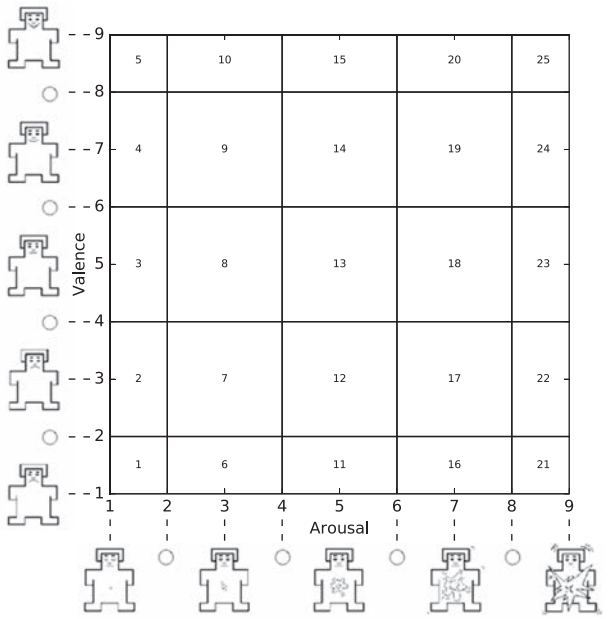

Fig. 2. Clustering for case 3 with the corresponding SAM choices and indices of areas in the AV space

2.2 Clustering of Synsets

We clustered the sound events based on the annotated emotion values, provided by the utilized datasets. In particular, we investigate the semantic similarity of sound events when the sound events are clustered according to their emotional annotation values in interpretable clusters, i.e., clusters that can be connected to either discrete emotions or particular areas in the AV field (e.g., high or low arousal) or choices from the emotion annotation tool that was employed during the emotional annotation. Therefore, we employed clusters of sound events that were formed by taking into account the values of their emotional annotations and not based on clustering algorithms.

For that reason five separate cases were considered, three for the IADS dataset and two for the BEADS, in order to investigate the semantic similarity of sound events according to different emotion-based clustering schemes. We first utilized a simple binary emotion-based clustering scheme for each separate affective dimension (case 1) and then we extended it by combining both affective dimensions (case 2). Then, we employed a SAM-based clustering due to the apparent relation that each SAM manikin has with the arousal and valence values (case 3). The remaining two cases (cases 4 and 5) were focused on the BEADS dataset and examined the semantic similarity of sound events when the latter are emotionally clustered according to the different emotion that they elicit when their source is moving towards the back of the listener, as presented in the published works related to BEADS dataset [4, 20].

In more detail, case 1 focused on investigating the semantic similarity of the sound events that elicit high or low values in each dimension, i.e., simulating a binary classification of emotion for each affective dimension. We employed two clusters for each affective state dimension (i.e., arousal and valence). One of the clusters had the sound events that were annotated with corresponding affective state value below 5 and the other with the remaining ones. This lead to two clusters for arousal and two for valence.

Case 2 was centered on examining the semantic similarity of general sounds that are in each quadrant of the AV space as can be seen in Fig. 1. This also lead to four clusters in total, each one containing the sound events that were annotated with values in the respective quadrant. Case 3 addressed the semantic similarity on AV space areas defined by SAM's values. An illustration of the resulting clustering is given in Fig. 2 along with the indices for the areas/clusters and the corresponding assignment of SAM values.

Case 4 focused on the semantic similarity between the two classes as defined in [4]. These two classes correspond to two different modes for the impact of spatial position of the source to affective states. The first class contains sound events that are rated with higher arousal and lower valence as they move towards the back of the listener's head, and the second consists of sound events that are rated with higher valence and lower arousal for the same movement of source. Two clusters were formed, each containing sound events belonging to the corresponding class, i.e., cluster 1 consisted of sound events appearing in class 1 and cluster 2 of sound events appeared in class 2. Finally, case 5 examined the semantic similarity of the sound events belonging to each aforementioned class but also according to angular transition, as employed in [4], where the class of each sound event is specified as the source moves toward the back of the listener's head. This movement is examined in the range of [0, 180] degrees and with a step of 45°. We clustered the sound events for each angular transition of the source, i.e.,

[2]http://wordnetweb.princeton.edu/perl/webwn.

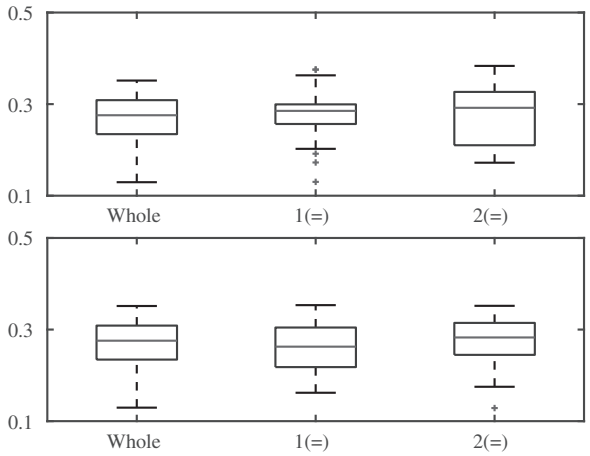

Fig. 3. The semantic similarity results for clustering case 1 in both clusters of (a) valence and (b) arousal.

0 to 45 degrees, 45 to 90 degrees, and so on up to 135 to 180 degrees, according to the class that they belong.

2.3 Similarity Measurement and Post-Processing

Semantic similarity was calculated for all synset pairs within a cluster and by utilizing the Wu-Palmer measure in conjunction with the matching similarity as described in [29]. For measuring the whole dataset similarity, as a measure of reference, all the synsets were considered to comprise a single cluster. In cases of sound events described by single synset only the Wu-Palmer measure was utilized. In the cases of sound events having two synsets, the matching similarity was employed [29]. In order to study the variations of the semantic similarity in each cluster, we averaged the similarity of each synset, i.e., we utilized the mean similarity of each synset with the others in the same cluster. Obtained values are presented in the form of boxplots in the next section.

3 RESULTS AND DISCUSSION

In all presented figures the semantic similarity for the whole dataset is also depicted as an indication of reference. Figs. 3 to 5 illustrate the obtained results for clustering cases 1 to 3, respectively, while Figs. 6 and 7 depict the semantic similarity results obtained for the clustering case 4 and 5, respectively. In these figures, statistical significance computed with the Wilcoxon rank-sum test [30] for higher and lower similarity values (compared to the similarity of the whole dataset), is indicated with (+) and (–) respectively, while statistically insignificant differences (compared to whole dataset) are marked with (=).

Quadrant indices on Fig. 4 follow counter-clockwise indexing, with number 1 assigned to the top-right quartile. Non referred areas in Fig. 5 are due to the lack of data according to Fig. 1. Angular transitions mentioned in Fig. 7 are according to [4].

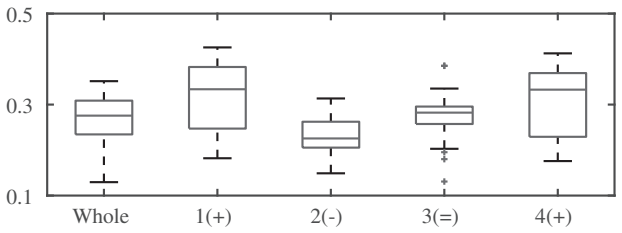

Fig. 4. The semantic similarity results for case 2.

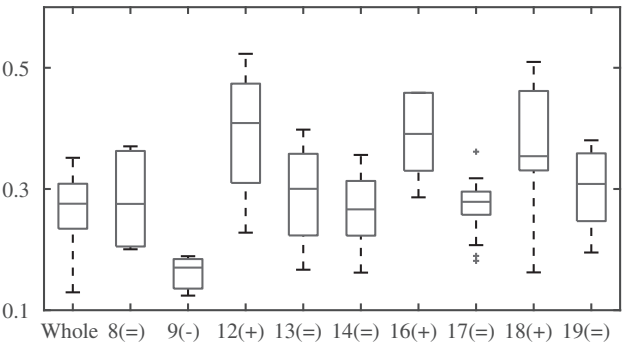

Fig. 5. The semantic similarity results for case 3.

The discussion of the results will be on the following to axis: (i) semantic similarity of sound events without spatial information (i.e., by employing the IADS dataset), and ii) semantic similarity combined with the spatial information of the source (i.e. by using the BEADS dataset).

For the former case, a close inspection on Figs. 3 to 5 reveals that in general the clustering of the sound events according to the elicited emotion does not exhibit a substantial increase on the semantic similarity with some exceptions. In particular, Fig. 3 shows that a clustering based on high or low affective state value (i.e., above or below the mean value of 5) does not have almost any result on the semantic similarity of the sound events. This indicates clearly that for classifying the emotional impact from sound events with the specific scheme, i.e., binary classification corresponding to high and low affective state values, the semantic content seems to have an insignificant effect. A fact that seems to be rather important since, on one hand, there are published works that employ such a grouping of sound events based on the arousal and valence annotations and, on the other, the binary classification can be considered as a first approach to the task of emotion recognition from sound events.

On the contrary, according to Figs. 4 and 5 there seems to be an increased effect of the semantic content (i.e., increased semantic similarity) to the elicited emotion as the clustering becomes more fine grained, in terms of areas on the AV space. But, again, in this case some of clusters do not portray such a behavior. Specifically, Figs. 2 and 5 reveal that the highest values for semantic similarity are observed for moderate to low valence and moderate to high arousal values. This observation can be also of high importance since it is stated in [19] that it is highly unlikely for one person to hear something that he does not like (low valence) and at the same time not feel aroused (low arousal). Additionally, valence seems to be considered as the most

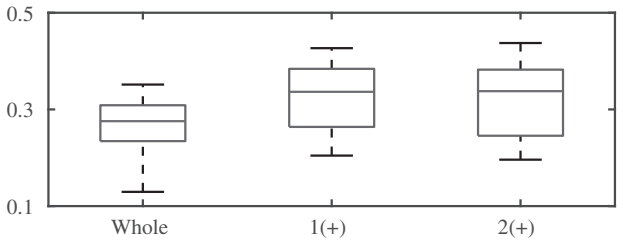

Fig. 6. The semantic similarity results for case 4.

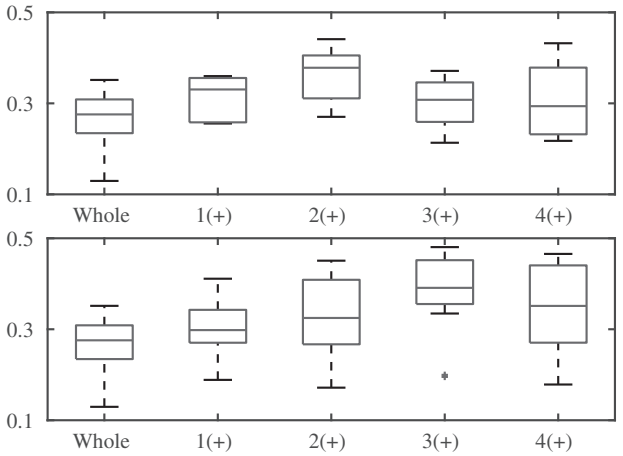

Fig. 7 The semantic similarity results for clustering case 1 in both clusters of (a) Class 1 and (b) Class 2.

difficult affective state to recognize [2], while with the presented results, the research focused on valence recognition can be benefitted by also employing semantic information. The p value obtained from Wilcoxon rank-sum test and for areas with indices 9, 12, 16, and 18 in Fig. 5 was below 0.01.

Regarding the combination of spatial information with semantic similarity, Figs. 6 and 7 show that, on one hand, the grouping in classes according to [4] leads to sets of sound events that exhibit higher semantic similarity in comparison to the mean one in and, on the other hand, there is an increased semantic similarity for the angular transitions that correspond to exact lateral positions with respect to the listener. For cases 4 and 5, all presented results had a p value below 0.01. Specifically, in Fig. 7 can be seen that for class 1, i.e., the sound events that elicit increased arousal and decreased valence as the source moves toward the side and the back of the listener, there is an almost double semantic similarity for the angular transition of 45 to 90 degrees. Furthermore, the same effect can be observed for class 2 and the angular transitions of 90 to 135 and 135 to 180 degrees. This aspect reveals that the sound events that tend to affect the elicited emotion to the listener according to the spatial location of their source tend to have semantic similar descriptions, indicating that the semantic content of sound events has an impact on the elicited emotion when combined with realistic spatial representation of the source (i.e., including the spatial information of the audio stimulus).

4 CONCLUSIONS

This work presented a first investigation on the potential impact of semantic content of sound events to the elicited emotion on the listener. For this approach two types of datasets were employed, one containing sound events without any spatial information and another consisting of binaural rendered sound events. Sound events were clustered according to their affective state annotations and the semantic similarity of the resulting clusters was measured by the utilization of WordNet and semantic similarity measures. Synsets used in the semantic similarity were created according to the textual description of each file in the utlized datasets. Results indicate that the semantic content has an impact mostly on the valence dimension and especially on mean to low valence values. In addition, the obtained results depicted that sound events that seem to have a systematic effect on the listener's emotion exhibit also an increased semantic similarity. This effect is more visible when the source of the audio stimulus is moving on the lateral and back areas of the listener's head.

The findings of the work at hand could initiate the research on the semantic similarity of sound events. Such research could reveal significant findings regarding the connection of sound events, their semantic content, and the elicited emotions.

5 REFERENCES

[1] K. Drossos, A. Floros, and N.-G. Kanellopoulos, "Affective Acoustic Ecology: Towards Emotionally Enhanced Sound Events," in *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound*, ser. AM '12. New York, NY, USA: ACM (2012), pp. 109–116.

[2] K. Drossos, A. Floros, and K.-L. Kermanidis, "Evaluating the Impact of Sound Events' Rhythm Characteristics to Listener's Valence," *J. Audio Eng. Soc.*, vol. 63, pp. 139–153 (2015 Mar.).

[3] M. Plewa and B. Kostek, "Music Mood Visualization Using Self-Organizing Maps," *Archives of Acoustics*, vol. 40, no. 4, pp. 513–525 (2015 Dec.).

[4] K. Drossos, A. Floros, A. Giannakoulopoulos, and N. Kanellopoulos, "Investigating the Impact of Sound Angular Position on the Listener Affective State," *Transactions on Affective Computing*, vol. 6, no. 1, pp. 27–42 (2015 Jan.).

[5] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, and S. Narayanan, "Automatic Recognition of Emotion Evoked by General Sound Events," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on* (2012 Mar.), pp. 341–344.

[6] M. Marcell, M. Malatanos, C. Leahy, and C. Comeaux, "Identifying, Rating, and Remembering Environmental Sound Events," *Behavior Research Methods*, vol. 39, no. 3, pp. 561–569 (2007).

[7] W. W. Gaver, "What in the World Do We Hear? An Ecological Approach to Auditory Event Perception," *Ecological Psychology*, vol. 5, pp. 1–29 (1993).

[8] E. Asutay, D. Västfjäll, A. Tajadura-Jiménez, A. Genell, P. Bergman, and M. Kleiner, "Emoacoustics: A

Study of the Psychoacoustical and Psychological Dimensions of Emotional Ssound Design," *J. Audio Eng. Soc.*, vol. 60, pp. 21–28 (2012 Jan./Feb.).

[9] P. Bergman, A. Sköld, D. Västfjäll, and N. Fransson, "Perceptual and Emotional Categorization of Sound," *J. Acous. Soc. Amer.*, vol. 126, no. 6, pp. 3156–3167 (2009 Dec.).

[10] P. Pertilä and J. Nikunen, "Distant Speech Separation Using Predicted Time–Frequency Masks from Spatial Features," *Speech Communication*, vol. 68, pp. 97–106 (2015).

[11] A. Diment, T. Heittola, and T. Virtanen, "Semi-Supervised Learning for Musical Instrument Recognition," in *21st European Signal Processing Conference 2013 (EUSIPCO 2013)* (2013 Sep.).

[12] J. Kauppinen, A. Klapuri, and T. Virtanen, "Music Self-Similarity Modeling Using Augmented Nonnegative Matrix Factorization of Block and Stripe Patterns," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on* (2013 Oct.), pp. 1–4.

[13] S. Krothapalli and S. Koolagudi, " Speech Emotion Recognition: A Review," in *Emotion Recognition Using Speech Features*, ser. SpringerBriefs in Electrical and Computer Engineering (Springer New York, 2013), pp. 15–34.

[14] P. N. Juslin and D. Västfjäll, "Emotional Responses to Music: The Need to Consider Underlying Mechanisms," *Behavioral and Brain Sciences*, vol. 31, pp. 559–575 (2008 Oct.).

[15] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music and Sound Have in Common," *Frontiers in Psychology*, vol. 4, no. 292 (2013).

[16] G. Lemaitre, O. Houix, N. Misdariis, and P. Susini, "Listener Expertise and Sound Identification Influence the Categorization of Environmental Sounds," *J. Experimental Psychology: Applied*, vol. 16, no. 1, pp. 16–32 (2010 Mar.).

[17] O. Houix, G. Lemaitre, N. Misdariis, P. Susini, and I. Urdapilleta, "A Lexical Analysis of Environmental Sound Categories," *J. Experimental Psychology: Applied*, vol. 18, no. 1, pp. 52–80 (2012 Mar.).

[18] G. Lemaitre and L. M. Heller, "Evidence for a Basic Level in a Taxonomy of Everyday Action Sounds," *Experimental Brain Research*, vol. 226, no. 2, pp. 253–264 (2013).

[19] M. M. Bradley and P. J. Lang, "International Affective Digitized Sounds (IADS): Stimuli, Instruction Manual and Affective Ratings," The Center for Research in Psychophysiology, University of Florida, Gainesville, Fl, Tech. Rep. B-2 (1999).

[20] K. Drossos, A. Floros, and A. Giannakoulopoulos, "BEADS: A Dataset of Binaural Emotionally Annotated Digital Sounds," in *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on* (2014 July), pp. 158–163.

[21] M. M. Bradley and P. J. Lang, "Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential," *J. Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59 (1994).

[22] Z. Wu and M. Palmer, "Verbs, Semantics and Lexical Selection," in *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ser. ACL '94 (Stroudsburg, PA, USA: Association for Computational Linguistics, 1994), pp. 133–138.

[23] G. A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41 (1995 Nov.).

[24] P. Ekman, "An Argument for Basic Emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200 (1992 Jan.).

[25] K. Drossos, R. Kotsakis, G. Kalliris, and A. Floros, "Sound Events and Emotions: Investigating the Relation of Rhythmic Characteristics and Arousal," in *Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on* (2013 July), pp. 1–6.

[26] H. Li, Y. Tian, B. Ye, and Q. Cai, "Comparison of Current Semantic Similarity Methods in WordNet," in *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, vol. 4 (2010 Oct.), pp. V4–408–V4–411.

[27] L. Meng, R. Huang, and J. Gu, "A Review of Semantic Similarity Measures in WordNet," *Interl. J. Hybrid Information Tech.*, vol. 6, no. 1, pp. 1–12 (2013 Jan.).

[28] [Online]. Available: http://www.nltk.org/howto/wordnet.html

[29] M. Rezaei and P. Fränti, *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings* (Springer Berlin Heidelberg, 2014), ch. "Matching Similarity for Keyword-Based Clustering," pp. 193–202.

[30] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83 (1945).

THE AUTHORS


Konstantinos Drossos


Maximos Kaliakatsos


Andreas Floros


Tuomas Virtanen

Konstantinos Drossos holds a B.Eng. equivalent on sound and musical instruments technology with distinction from Technological Educational Institute of Ionian Islands (Kefalonia, Greece), an M.Sc. from I.S.V.R., University of Southampton (Southampton, UK), and in 2015 he received his Ph.D. on emotional information retrieval from sound events from the Audiovisual Arts Dept, Ionian University (Corfu, Greece). Currently he is a post doctoral researcher at the Audio Research Group, Dept. of Signal Processing, Tampere University of Technology,Tampere, Finland. His main research interests are emotion recognition from sound events, sound events recognition, audio perception, and audio interfaces. He has also worked as a freelance acoustic consultant, audio programmer, Adjunct Lecturer at the dept. of Sound and Musical Instruments Technology of the Technological Educational Institute of Ionian Islands, and as researcher in various research and development projects. Dr. Drossos is a member of the Institute of Electrical and Electronics Engineers and Hellenic Institute of Acoustics.

•

Maximos Kaliakatsos studied mathematics at the University of Patras, Greece, where he also received a Masters Degree in computational intelligence (CI) and a Ph.D. on applications of CI on music and sound. Among his research interests is the utilization of evolutionary processes for rhythm, harmony, and melody generation as well as computational approaches to extract information out of music and audio. He is currently working as a research fellow at the School of Music studies in the Aristotle University of Thessaloniki and he is also a member of the Cognitive and Computational Musicology (CCM) group at this department. His current research interest is the application of artificial intelligence methods and conceptual blending in automated creative melodic harmonization.

•

Andreas Floros was born in Drama, Greece, in 1973. In 1996 he received his engineering degree from the department of electrical and computer engineering, University of Patras, and in 2001 his Ph.D. degree from the same department. His research was mainly focused on digital audio signal processing and conversion techniques for all-digital power amplification methods. He was also involved in research in the area of acoustics. In 2001, he joined the semiconductors industry, where he worked in projects in the area of digital audio delivery over PANs and WLANs, Quality-of-Service, mesh networking, wireless VoIP technologies, and lately with audio encoding and compression implementations in embedded processors. During 2003–2005 he was a member of a number of IEEE Tasks Groups (such as the 802.11e, .11k and .11s) with voting rights. For a period of three years (2005–2008), he was an adjunct professor at the department of informatics, Ionian University. During this period of time he also taught at the postgraduate (M.Sc.) degree "Arts and Technologies of Sound" organized by the dept. of Music Studies, Ionian University. On January 2008, he was appointed in the position of Assistant Professor at the department of Audiovisual Arts, Ionian University. Today, he is an Associate Professor and the head of the above department. His current research interests focus on analysis, processing and conversion of digital audio signals, intelligent digital audio effects and sound synthesis, creative intelligence, audio-only games, auditory interfaces and displays, augmented reality audio foundations, and applications as well as the investigation of the impact of sound events to human emotions. Dr. Floros is a member of the Audio Engineering Society (currently serving as the Secretary of the AES Greek Section), while he actively participates in the AES Technical Committee on Network Audio Systems and Audio for Games.

•

Tuomas Virtanen is an Academy Research Fellow and Associate Professor (tenure track) at Department of Signal Processing, Tampere University of Technology (TUT), Finland, where he is leading the Audio Research Group. He received the M.Sc. and Doctor of Science degrees in information technology from TUT in 2001 and 2006, respectively. He has also been working as a research associate at Cambridge University Engineering Department, UK. He is known for his pioneering work on single-channel sound source separation using non-negative matrix factorization based techniques and their application to noise-robust speech recognition, music content analysis, and audio event detection. In addition to the above topics, his research interests include content analysis of audio signals in general and machine learning. He has authored more than 100 scientific publications on the above topics, which have been cited more than 3000 times. He has received the IEEE Signal Processing Society 2012 best paper award for his article "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria" as well as three other best paper awards. He is an IEEE Senior Member and recipient of the ERC 2014 Starting Grant, and a member of the Audio and Acoustic Signal Processing Technical Committee of IEEE Signal Processing Society.