BEADS: A Dataset of Binaural Emotionally Annotated Digital Sounds

Konstantinos Drossos
Audiovisual Signal Processing Lab.
Dept. of Audiovisual Arts
Ionian University
Corfu, Greece
Email: kdrosos@ionio.gr

Andreas Floros
Audiovisual Signal Processing Lab.
Dept. of Audiovisual Arts
Ionian University
Corfu, Greece
Email: floros@ionio.gr

Andreas Giannakoulopoulos
Interactive Arts Lab.
Dept. of Audiovisual Arts
Ionian University
Corfu, Greece
Email: agiannak@ionio.gr

***Abstract*—Emotion recognition from generalized sounds is an interdisciplinary and emerging field of research. A vital requirement for this kind of investigations is the availability of ground truth datasets. Currently, there are 2 freely available datasets of emotionally annotated sounds, which, however, do not include sound evenets (SEs) with manifestation of the spatial location of the source. The latter is an inherent natural component of SEs, since all sound sources in real-world conditions are physically located and perceived somewhere in the listener's surrounding space. In this work we present a novel emotionally annotated sounds dataset consisting of 32 SEs that are spatially rendered using appropriate binaural processing. All SEs in the dataset are available in 5 spatial positions corresponding to source/receiver angles equal to 0, 45, 90, 135 and 180 degrees. We have used the IADS dataset as the initial collection of SEs prior to binaural processing. The annotation measures obtained for the novel binaural dataset demonstrate a significant accordance with the existing IADS dataset, while small ratings declinations illustrate a perceptual adaptation imposed by the more realistic SEs spatial representation.**

I. Introduction

Sound is ubiquitous and we receive stimuli to our auditory organ almost constantly. The majority of these stimuli are non–linguistic and non–musical sounds demonstrating a non–organized form in terms of melody and rhythm, in contrary to music [1]. Such prompts are termed general sounds or sound events (SEs) [2]. They emanate from various sources, e.g. human activities, objects' interactions etc., communicate various information, like the nature of the sound producing mechanism or the instantaneous spatial location of the source, and construct our acoustic environment [3], which is the main field of research for the Acoustic Ecology (AE) discipline [4]. It is widely known that SEs elicit emotions to the acoustic receiver [5]. This fact has led to a proliferation of research towards the emotion recognition from general sounds [6] and to the recent introduction of the Affective Acoustic Ecology concept [1]. The latter has as its key element the SE itself and expands the AE definition by regarding also the emotion conveyed and elicited to the listener by the audio environment [1].

Although emotion recognition from music signals represents a well-investigated concept that has led to numerous systematic conclusions that outline the indubitable relation between music and human emotions, the field of emotion recognition from generalized sonic content is rather recent and unfolding [7]. Possible applications span from audio enhancement for entertainment purposes (e.g. video games) to emotional augmentation of human immersion for artificial environments (e.g. augmented or virtual reality applications) [1]. The limited, existing works that investigate emotion recognition from SEs use emotionally pre-annotated sounds datasets with accuracy results reaching up to 88% [6]. Since this is a recent field of research, the available audio data sets are of paramount importance. According to the authors' knowledge, only two such emotionally annotated SEs are reported: the International Affective Digital Sounds (IADS) dataset [5] and the Emotional Sound Database introduced in [8].

The above emotionally annotated sound datasets were employed by research that considered a rather limited definition and parameterization of the SEs under test. For example, it was assumed that a sound event is modeled by a single channel sound waveform (i.e. a monophonic recording), thus excluding any kind of spatial information that is inherent into a SE formulation. This information is perceptually important, since the spatial position of the sound source is naturally communicated to the listener [3]. Nevertheless, an additional work [9] did considered this extended SE definition and particularly investigated the impact of the sound source spatial position to the elicited "Fear" using a rather vague approach: by placing the source at the front and the back side of the listener, thus providing early indications regarding the important role of the sonic environment spatial characteristics on a primitive emotion. The systematic and extended exploration of this topic, in both terms of sound source spatial representation accuracy and targeted emotions, brings forward the lack of a dataset with emotionally annotated SEs which can manifest the sources' spatial location.

Towards fulfilling this gap, in the work at hand we present a novel emotionally annotated sound dataset with spatially positioned SEs. In particular, we have used a subset of the original IADS dataset [5] for obtaining the raw SEs waveform. Then, we have used binaural rendering in order to infuse the desired spatial information. The latter was limited in the horizontal plane only, thus allowing the employment of the new dataset by future investigations with particular focus on two dimensions (2D). In typical sound reproduction setups, spatial positioning of the source can be achieved under numerous approaches. From these, and for the purposes of this work, we selected binaural technology, which provides a robust means for delivering accurate three dimensional sound field

representations using only two discrete audio channels, at the expense of optimized reproduction using headphones. It is widely–known that binaural technology is based on filtering ideally anechoic sound recordings with filters that model the human body (i.e. the head and the upper part of the shoulders) and the outer ear influence on the received sound signal. These filters are therefore termed as *Head Related Transfer Functions* (HRTFs) and are obviously defined as a function of the horizontal and vertical angle appeared between the sound source and the listener's head [10].

In particular, we employed the KEMAR HRTF library [10] for binaural rendering in the horizontal plane. All sources were regarded as facing towards the listener, being located in a free field. 5 different angular positions around the listener were defined, equal to 0^o, 45^o, 90^o, 135^o, 180^o. Due the symmetrical properties of the HRTFs in the horizontal plane, the above angular resolution sufficiently covers all the cases where the sound source is located inside, in the lateral positions and outside of the listener's view, providing an adequate extension of the spatial coverage to existing data sets.

Furthermore, emotional annotation was performed using the Self Assessment Manikin (SAM) method [11]. All subjective experiments were performed through a custom developed web platform, a fact that introduced a number of advantages that will be analyzed next in the paper. The obtained subjective ratings were compared to the original IADS ones, showing a close match between them, but more importantly, leading to an indication that the existing emotionally annotated datasets suffer from inaccuracies imposed by disregarding important SEs parameters always present in real-world sonic environments. Hence, one can assume that the presented dataset is likely to boost the research focus on the impact of generalized sound events to the elicited affective state of the listener.

The rest of the paper is organized as following: In Section II we present a concise overview of the existing emotionally annotated sounds datasets and the employed techniques for emotion annotation. Section III provides a description of the details of the presented Binaural Emotionally Annotated Digital Sounds (BEADS) dataset. Next, Section IV presents a summary of the received annotations, which are further discussed and compared to existing ones in Section V. Finally, Section VI concludes the work and highlights some particular issues that can be considered for future enhancements.

II. Existing emotionally annotated sound datasets

In general, emotion recognition from audio signals is modeled as a pattern matching task [7]. Therefore, a typical work–flow of this process consists of a training stage, where the ground truth dataset is initially analyzed. Then, a model based on this analysis is created, and a testing phase is following, for evaluating the developed model. A graphical representation of this work-flow is illustrated in Figure 1. Clearly, one can distinguish two key elements in the above process: a) the required datasets in the training and testing phases, and b) the subjective annotation of these datasets. This Section focuses particularly on these two elements and presents existing emotionally annotated SE sets along with the annotation method followed for their realization.

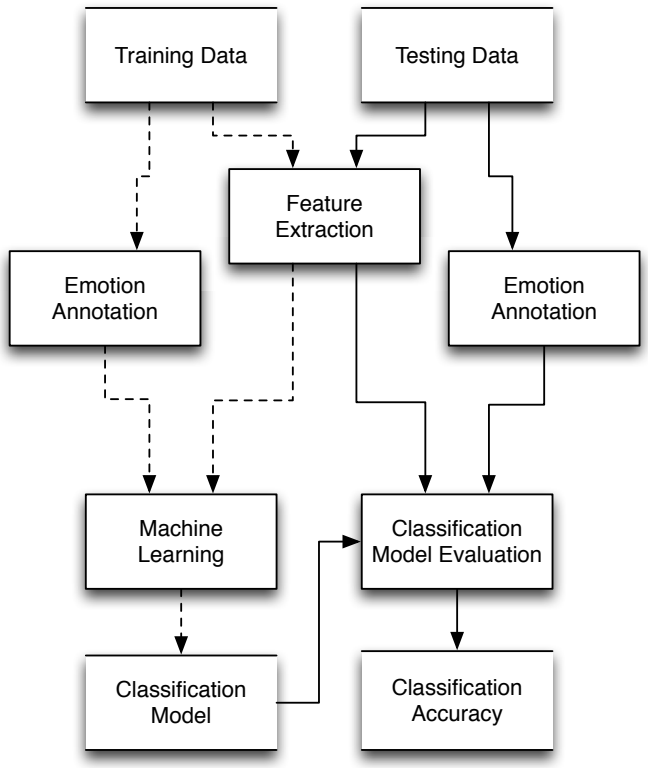

Fig. 1. A typical work-flow of the emotion recognition from sound process. The dashed line indicates the training stage and the solid the testing one.

A. The IADS dataset

The International Affective Digital Sounds (IADS) library was originally introduced in [5] and is available online [12]. It consists of 167 sounds, all having a time length of 6 seconds and sampling frequencies ranging from 8 to 44.1kHz. Sound sources utilized in the IADS dataset are both single ones, e.g. a phone ringing, and ambient sounds, e.g. sounds in a tropical forest. Despite this twofold character of the sound sources, all sound waveforms in the dataset are monophonic, thus containing no information that can be conveyed to the listener regarding the sound source position. Their semantic content also varies greatly and includes a wide range of events that can occur in everyday life, such as:

- Animal sounds: dogs barking, growls, chickens and rooster, rattle snake noise, robin, bees, pigs and cows/cattle
- Human activities: erotical actions and interjections, laughings, vomiting, sneezing, wheezing, heartbeat, cries from babies and grown ups, yawning, fights, screams, sobbing, crowd noise, writing, countdown, talking, clapping, chewing, whistling, hiccup, giggling and singing
- Sounds created from objects or interaction with objects: carousel, music box, video game, gunshots, type writter, polaroid, lawnmower, doorbell, car horns, engine failure, crashes and explosions, fan, phone sounds (ringing and busy signal), dentist drill, buzzer, sirens, slot machines, bottle opening, helicopters, shovel and

jackhammer

- Ambient sounds: rain, tropical forest, fight, air raid sirens, wind, sound of sink, water flowing, sound from people gathering (e.g. in a bar) and night sounds
- Musical SEs; different musical excerpts, choir and musical instruments

The annotation of the IADS dataset was performed based on the SAM rating principles and by considering the arousal, valence and dominance affective model. The latter represents an emotional model widely accepted as valid by the music emotion recognition community [13], [14]. During the experimental annotation sequence, sound reproduction was performed using a laptop computer and a set of loudspeakers. No specific sound processing tasks are reported by the authors, apart from anti-clipping protection. Ultimately, all sounds were annotated for all 3 affective dimensions (i.e. arousal, valence and dominance). Each sound was rated by approximately 100 human subjects, with half of them being females.

B. The Emotional Sound Database

Recently, another dataset with emotionally annotated general sounds was presented in [8] named Emotional Sound Database. It can be also found online [15]. The SEs for constructing it were retrieved from the online sound library FindSounds [16]. Specifically, it consists of 360 ambient and non–ambient SEs, with variable sampling frequencies and durations. SEs in FindSounds database are clustered in 16 categories with respect to their semantic content. Emotional Sound Database follows the same categorization.

The annotation of the Emotional Sound Database was performed with the participation of only 4 subjects, who provided their ratings for both the arousal and valence affective dimensions. Additionally, the evaluator weighter estimator [17] was used in order to increase the robustness of the obtained annotations. No spatial information on the SEs was also included.

III. The BEADS Dataset

A common restriction of both the above emotionally annotated datasets is that they do not incorporate SEs representations that are compatible with the Affective Acoustic Ecology scope, i.e. they do not consider the affective impact of the spatial properties of sound. In this paper we introduce the BEADS dataset that can lift the above limitations by offering to the research community a standard sound corpus with binaural emotionally annotated digital sounds.

A. Sound corpus derivation

As it was mentioned earlier, the BEADS dataset is based on the IADS [5] sound corpus, while sound spatial positioning around the listener is performed using binaural processing. Five different positions are considered, with horizontal angles equal to $\theta(k)$=$\{0^o,\ 45^o,\ 90^o,\ 135^o,\ 180^o\}$, with $k \in [1, 5]$ being the corresponding *spatial index*. More specifically, if we assume that the IADS dataset is denoted as $s(i)$ (with $i \in [1,\ 167]$), then the binaural versions of the original sounds are calculated as:

$$s_b(i,k) = s(i) * h(k) \tag{1}$$

where $*$ indicates convolution and $h(k)$ denotes the employed KEMAR HRTFs [10] for the k-th angle. Thus, a total amount of 835 spatialized sound events were created, which were then normalized to -4.44 dB relative to Full Scale (dBFS) in order to minimize the possibility of clipping occurrences caused by high reproduction gain adjustments performed by the human annotators. The derived binaural sound set was finally organized in terms of 5 angular subsets $S_a(k)$, each containing all 167 $s_b(i,k)$ binaural waveforms, i.e. $S_a(k) = \{s_b(1,k),\ s_b(2,k),\ \ldots,\ s_b(167,k)\}$. The $S_a(k)$ subsets were stored in an SQL–compliant database organized in tuples $t_b(i,k)$ formed as:

$$t_b(i,k) = (s_b(i,k), c(i,k)) \tag{2}$$

where $c(i,k)$ is a variable employed in order to count the number of times that a particular $s_b(i,k)$ binaural signal was selected for reproduction in subsequent listening tests. The exact purpose of this counter is explained thoroughly in the next Section.

B. Sound corpus annotation

A series of subjective evaluation experiments was carried out in order to obtain the affective state annotations. Following the IADS dataset case, the annotations were performed using an extension of the original SAM [11] method that employs extra intermediate states for both the arousal and valence dimensions [5]. We selected a $2D$ rather than a $3D$ affective model, since it is reported that the first one imposes decreased complexity [18]. Moreover, the intermediate SAM states were applied in order to introduce a coherence to the obtained affective state ratings between the BEADS and IADS datasets [5] .

All subjective rating tests were accomplished using a web platform[1] developed for the purposes of this work. Invitations for participation were sent via batch e-mails targeted to members of specific mailing lists that are active in the areas of audio engineering and psychoacoustics, such as the auditory mailing list [19]. Invitations to specific individuals were also sent using their institutional e-mail addresses. The employment of a web-based environment for realizing the subjective ratings provided a number of advantages compared to legacy listening tests implementation approaches such as:

1) The world-wide distribution of the origin of the participants, which at a large extent was verified through their institutional e-mail addresses. It turned out that subjects from at least three continents (Europe, north and south America and Asia) responded and participated. This fact reduced the potential correlation of the annotations with the semantic content affected by transcontinental, national and local cultural differences.
2) Simultaneous participation without any personal or geographical time restrictions: The participants could have access to the experiment's web environment concurrently at their own personal convenience.
3) It is very likely that each participant carried out the experiment when he had the time and mood to do so. Hence, one can assume that the participants felt

[1] http://audemo.eu/en

comfortable and paid the appropriate attention during their rating session. Consequently, the results would not be affected by any emotional conditions raised and conveyed by a laboratory environment and a strict scheduled plan for participation.

Nevertheless, the web–based experimental execution also had some drawbacks and risks. For example, headphone equalization for accurate binaural reproduction was not possible. More importantly, although the subjects were strongly requested to use headphones, one can consider that some of them did not. Generally speaking, such procedural issues are subject to the trust that should be ascribed to the subjects participating in an annotation session (or in a subjective test in general). For example, the aforementioned possibility of not correctly following the guidelines provided could be also applied in a locally–executed, fully-controlled experiment by arguing the participants' answering truthfulness. Finally, specific caution was taken for avoiding multiple participations of the same subjects. Towards this aim, a login by e–mail secure mechanism was applied, allowing only one login session per individual e–mail address. However, since this e–mail based subject identification mechanism can be easily breached and the same person could make a different e–mail account and retake the experiment, we followed an additional soft control approach. Upon completing the experiment each participant could repeat the process but the results were not stored to the database. There was no indication to the participant for this difference and thus the need for pretending a different person in order to retake the experiment was reduced.

The actual experimental sequence consisted of two parts. In the first one, a detailed description of the subjective tests was provided to all participants. Additionally, the latter followed specific instructions in order to appropriately adjust the reproduction level prior to the listening tests. In the second part, each subject had to sequentially listen to 15 sounds from the binaural sound corpus and rate them using the SAM method.

In particular, after the successful login of each subject, a set of informative messages was appeared regarding a) the necessity for using headphones during the test and b) the adjustment of the reproduction volume along with a selection for the reproduction of a 0dBFS 1kHz pure tone, served as the reference sound for the level adjustment calibration. Under this task, each participant had to adjust the sound reproduction level to the maximum that he/she feels comfortable, provided that the sound was not perceptually distorted. The 1kHz frequency was selected due to the fact that (a) the human auditory system does not introduce any weighting in this frequency band and (b) it is most likely that all electroacoustic transducers have a frequency response gain equal to 0 dB for this particular frequency. Then, the subject was asked to retain the above level adjustment constant until the end of the experimental session. This step was necessary, since each subject used his own headphone equipment and sound card interface, rendering impossible to apply a unique reproduction gain definition process. Next, an on–line video was shown, demonstrating the exact experimental process and the sequence of actions that would be followed during the experiment. It was also clearly stated that there was no correct or wrong answer to the subjective ratings.

The second part of the experiment was initiated by the participant. At the beginning, a playlist containing 15 $t_b(i,k)$ tuples was automatically created. These tuples were selected randomly using standard SQL commands under the following conditions:

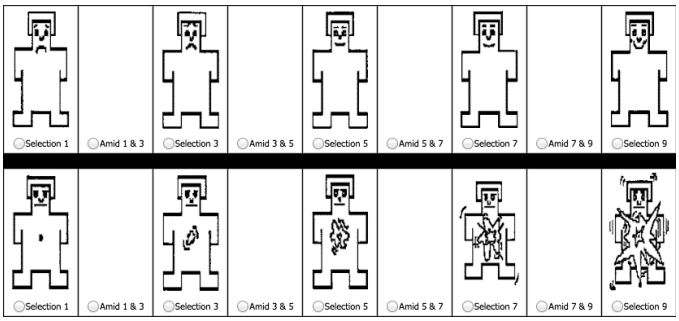

Fig. 2. The available affective annotation choices

$$T_b(i,k) = \min_{c(i,k)}(t_b(i,k)) \quad (3)$$

$$N(T_b(i,k)) = 3 \quad (4)$$

where $T_b(i,k)$ is the selected tuple and $N(T_b(i,k)$ denotes the amount of selected tuples for a particular angular position k. For all selected $T_b(i,k)$ tuples, the $c(i,k)$ value was then increased by one. This scheme ensured the uniform random selection of the available tuples.

Upon the creation of the binaural playlist, the corresponding binaural sounds were reproduced one by one and the participant was asked to provide his subjective valence and arousal ratings in terms of the appeared SAM drawings (see Figure 2). The assigned annotation values were stored within the sound corpus database. For reasons of clarity, the overall subjective evaluation process that was described here is graphically analyzed in Figure 3.

C. Sound corpus post processing

Due to the remote execution of the experiment, it was observed that some of the participants did not complete it. Thus, their ratings had to be excluded from the final results derivation. This fact additionally resulted into missing ratings for some $s_b(i,k)$. Hence, during this stage, the following tasks were performed: (a) all ratings for all sounds were arithmetically extracted from the ratings database (b) any non–valid participation was deleted and (c) $s_b(i,k)$ that were not rated for all possible k values were also excluded. This process led to a total of 32 SEs that were annotated for all k values or, effectively, to a 160 total set of binaural sounds. For reason of presentation simplicity, these SEs that were finally included in the BEADS sound corpus are denoted as $s_b(i',k)$. The new i' index is expressed as:

$$i = i' + f(i) \quad (5)$$

where $f(i)$ is an integer representing the total number of the sounds excluded up to the i-th IADS sound. Table I contains the complete list of the incorporated $s_b(i',k)$ SEs, along with the corresponding i and i' values and the respective semantic content. It turned out that each $s_b(i',k)$ SE received an average of 9 annotations for both arousal and valence dimensions. The

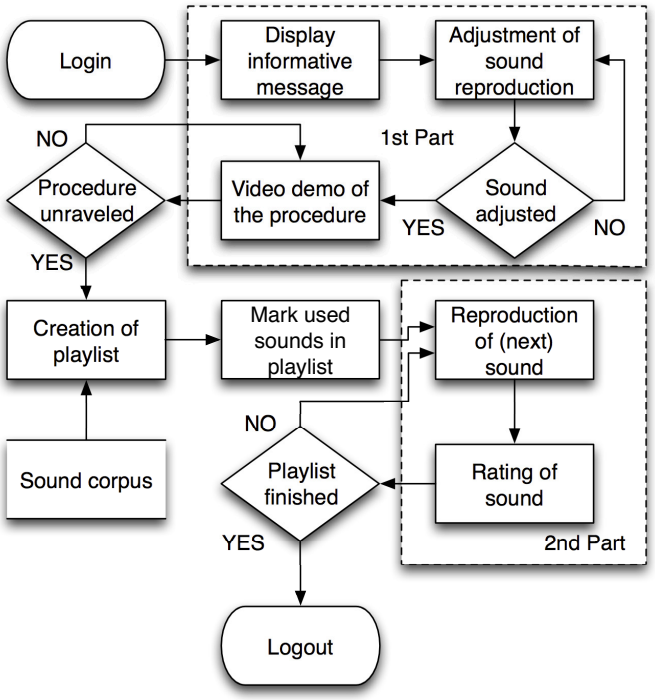

Fig. 3. A graphical representation of the listening tests organization

TABLE I. DETAILS OF THE SES INCLUDED IN THE FINAL BEADS DATA SET.

i'/i index	Semantic Content	i'/i index	Semantic Content
1/5	Dog	17/99	Rain
2/16	Chickens	18/106	Countdown
3/18	Rattle snake	19/107	Car horns
4/20	Robin	20/112	Wind
5/21	Tropical	21/113	Plane crash
6/26	Erotic female	22/114	Engine failure
7/35	Male laugh	23/115	Bike wreck
8/38	Laughing	24/122	Explosion
9/42	Couple sneeze	25/130	Phone
10/50	Vomit	26/133	Clock
11/54	Whistling	27/135	Cuckoo
12/62	Woman crying	28/142	Slot machine
13/66	Victim	29/146	Walking
14/80	Type writter	30/157	Harp
15/85	Writing	31/159	Bach
16/96	Sink	32/160	Choir

mean annotations values for these affective dimensions were calculated and they were finally stored in the BEADS dataset. The complete BEADS sound corpus is available online [20].

IV. ANNOTATION RESULTS

In order to estimate the accuracy of the BEADS annotations, we compared them with the IADS ones. Since the latter SEs are monophonic, their reproduction through a normal couple of loudspeakers (which is the playback setup followed for obtaining the IADS subjective scores) results into the placement of the virtual SE source exactly in front of the listener (we assume that the receiver is positioned within the sweet spot area defined for stereo reproduction). Hence, the above comparison can be performed by considering the BEADS SEs for $k = 1$. The scatter plot of the respective annotations in the valence / arousal space is illustrated in Figure 4.

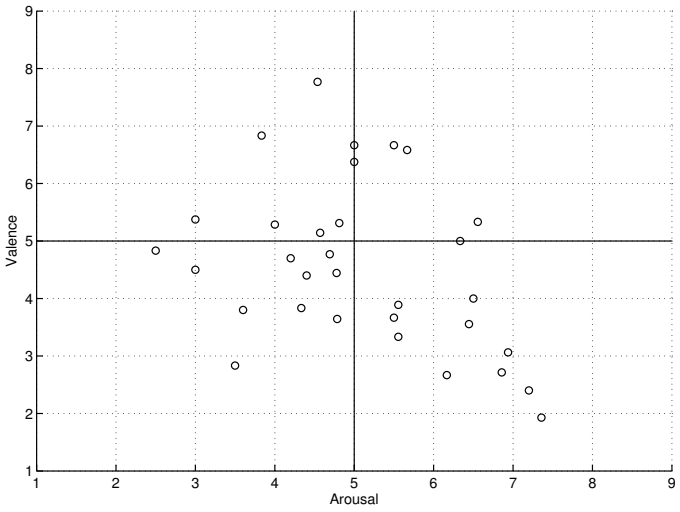

Fig. 4. Arousal and valence mean values for the $s_b(i', 1)$ angular subset ($\theta(k)$=0°)

TABLE II. DIFFERENCE OF THE MEAN AROUSAL AND VALENCE ANNOTATION RATINGS OF THE BEADS AND THE IADS DATASET

i'	Arousal difference ($D_a(i')$)	Valence difference ($D_v(i')$)	i'	Arousal difference ($D_a(i')$)	Valence difference ($D_v(i')$)
1	2.25	1.67	17	-0.27	1.14
2	0.77	0.35	18	0.22	1.46
3	2.20	-0.89	19	0.22	-0.37
4	-1.03	0.45	20	-0.10	0.65
5	1.68	-1.60	21	1.37	-0.59
6	1.64	0.12	22	1.99	-0.68
7	0.05	-0.11	23	1.11	-0.54
8	0.88	0.01	24	-0.33	0.31
9	1.69	1.03	25	0.99	1.60
10	-0.61	1.78	26	-1.23	0.70
11	-0.58	0.79	27	-0.20	0.71
12	0.70	-1.08	28	0.00	1.99
13	0.52	-0.25	29	2.47	0.00
14	0.10	0.24	30	0.36	2.07
15	0.47	0.12	31	-0.05	1.03
16	1.23	1.10	32	-1.14	1.76
Mean arousal:		-0.54	Mean valence:		0.47
σ of arousal:		+1.02	σ of valence:		0.95
Absolute mean of arousal:		0.89	Absolute mean of valence:		0.85
Absolute σ of arousal:		0.72	Absolute σ of valence:		0.62

The rating differences between the IADS and BEADS cases can be calculated as:

$$D_a(i') = A_{IADS}(i) - A_{BEADS}(i', 1) \quad (6)$$
$$D_v(i') = V_{IADS}(i) - V_{BEADS}(i', 1) \quad (7)$$

where $A_{IADS}(i)$ and $V_{IADS}(i)$ are the arousal and valence IADS ratings, and $A_{BEADS}(i', 1)$ and $V_{BEADS}(i', 1)$ are the mean arousal and valence scores assigned to the BEADS $s_b(i', 1)$ sound event respectively. These difference values are shown in Table II.

V. DISCUSSION

Focusing on Table II it can be observed that the maximum arousal difference equals to 2.47, while for the valence dimension the corresponding value is 2.07. On the other hand, the absolute mean difference values are 0.89 and 0.85 respectively.

The aforementioned differences indicate that the mean variation in the emotional annotations are marginally equal to one intermediate SAM state. In addition, from the non–absolute mean difference values it can be inferred that the BEADS SEs for $\theta(k)$=0^{o} elicit greater arousal and lower valence. Recapitulating the above arithmetic trends, it is obvious that the BEADS and IADS arousal/valence ratings exhibited a relative coherence, with the spatial information infused on the IADS dataset causing only slight variations on the listeners' affective components.

These variations though can be considered as the appropriate adjustments of the IADS ratings towards a more realistic SEs representation that encapsulates the inherently present spatial sound information: in [9] it has been already reported that a sound source with less spread tends to elicit less activation to the listener. Since most of the sources in the BEADS dataset are non–ambient sounds placed on a specific position in the auditory horizon of the listener, their spread is limited and the above adjustments are perfectly justified. Nevertheless, the aforementioned adjustments can not be considered valid for all sound source cases since there are 2 particular SEs (with i' equal to 5 and 17) which exhibit rather ambient characteristics, due to their specific semantic content. For these SEs the annotations of the IADS dataset can be regarded as more close to a real-world scenario.

VI. Conclusions

Sound emotion recognition represents an emerging field of research. There are several important issues that are not yet investigated, such as the relation of the inherent characteristics of generalized sound events with the affective state of the human listener. Such investigations should extend beyond the legacy analyses performed that typically considers generalized sound as simple sonic waveforms. Hence, ground-truth datasets should be available that contain the complete extent of information that it is naturally conveyed to the human listeners during the occurrence of a SE.

Towards this research potential, this work aims to resolve the lack of an emotionally annotated sound corpus that encapsulates the information related to the spatial position of the sound source relatively to the axis of the acoustic receiver and introduces the Binaural Emotionally Annotated Digital Sounds (BEADS) dataset. This collection of sounds includes binaural replicas of 32 SEs that were obtained from the existing IDAS sonic set, calculated for a range of spatial positions that cover the complete horizontal plane around the listener. These binaural sounds were annotated by a large number of human subjects using the valence / arousal affective model, providing a robust set of subjective evaluations that can be used in relative future research. The annotation process was performed through a multimedia web platform, allowing for the participation of subjects originating from multiple countries and cultures and further eliminating aggravated conditions related to user convenience and fatigue.

A limited set of preliminary results has shown that the obtained annotations demonstrate a significant coherence with the subjective scores included in the IADS dataset. Small variations however in the valence / arousal affective components do exist, that can be considered as the necessary adjustment required for exceeding the absence of any kind of sound spatial information of the original IADS dataset. Future extensions of the BEADS dataset can be further applied, mainly in terms of including additional annotated SEs, thus extending the variety of the included semantic content.

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*. ACM, 2012, pp. 109–116.

[2] 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. [Online]. Available: http://dx.doi.org/10.3758/BF03193026

[3] W. W. Gaver, "What in the world do we hear? an ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[4] B. Y. Newman, "And now, acoustic ecology," *Optometry - Journal of the American Optometric Association*, vol. 76, no. 11, pp. 629–631, Nov. 2013.

[5] M. M. Bradley and P. J. Lang, "The international affective digitized sounds (2nd edition; iads-2): Affective ratings of sounds and instruction manual," NIMH Center for the Study of Emotion and Attention, Gainesville, Fl, Tech. Rep. B–3, 2007.

[6] 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*, July 2013, pp. 1–6.

[7] 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, May 2013.

[8] 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*, March 2012, pp. 341–344.

[9] I. Ekman and R. Kajastila, "Localization cues affect emotional judgments - results from a user study on scary sound," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, Feb 2009.

[10] B. Gardner and K. Martin, "Hrtf measurements of a kemar dummy-head microphone," MIT Media Lab Perceptual Computing, Tech. Rep. 280, 1994.

[11] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[12] The center for the study of emotion and attention. [Online]. Available: http://csea.phhp.ufl.edu/media/iadsmessage.html

[13] M. M. Bradley and P. J. Lang, "Affective reactions to acoustic stimuli," *Psychophysiology*, vol. 37, no. 2, pp. 204–215, 2000.

[14] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273 – 294, 1977.

[15] openaudio.eu. [Online]. Available: http://www.openaudio.eu

[16] Findsounds - search the web for sounds. [Online]. Available: www.findsounds.com

[17] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10-11, pp. 787–800, Oct. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2007.01.010

[18] C. Stickel *et al.*, "Emotion detection: Application of the valence arousal space for rapid biological usability testing to enhance universal access," in *Universal Access in Human-Computer Interaction. Addressing Diversity*, ser. Lecture Notes in Computer Science, C. Stephanidis, Ed. Springer Berlin Heidelberg, 2009, vol. 5614, pp. 615–624.

[19] Auditory list home page. [Online]. Available: http://www.auditory.org

[20] Epoasi - research - publications. [Online]. Available: http://epoasi.eu/en/research/publications/research-material/2014