EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

AudioPairBank: towards a large-scale tag-pair-based audio content analysis

Sebastian Säger[1], Benjamin Elizalde[2*], Damian Borth[1], Christian Schulze[1], Bhiksha Raj[2] and Ian Lane[2]

Abstract

Recently, sound recognition has been used to identify sounds, such as the sound of a car, or a river. However, sounds have nuances that may be better described by adjective-noun pairs such as "slow car" and verb-noun pairs such as "flying insects," which are underexplored. Therefore, this work investigates the relationship between audio content and both adjective-noun pairs and verb-noun pairs. Due to the lack of datasets with these kinds of annotations, we collected and processed the AudioPairBank corpus consisting of a combined total of 1123 pairs and over 33,000 audio files. In this paper, we include previously unavailable documentation of the challenges and implications of collecting audio recordings with these types of labels. We have also shown the degree of correlation between the audio content and the labels through classification experiments, which yielded 70% accuracy. The results and study in this paper encourage further exploration of the nuances in sounds and are meant to complement similar research performed on images and text in multimedia analysis.

Keywords: Sound event database, Audio content analysis, Machine learning, Signal processing

1 Introduction

The ability to interpret sounds is essential to how humans perceive and interact with the world. Sounds are captured in recordings—mainly in videos—and the acoustic information captured is exploited in a number of applications. The dominant application is multimedia video content analysis, where audio is combined with images and text [1–4] to index, search, and retrieve videos. Another application is human-computer interaction and robotics [5, 6] where sounds (e.g., laughing, clapping) complement speech as non-verbal communication. This is particularly useful for visually impaired and blind computer users [7]. Newer applications include smart homes, where sounds such as *water tap running* are detected [8] to prevent waste, and smart cities [9–11], where acoustic pollution is defined by a set of sounds. All of these applications rely on automatic recognition of "audio concepts'—relevant acoustic phenomena that present as identifiable sound signatures within recordings.

Sound recognition research in the past few years has been advanced by competitions and standard datasets. From 2010 to 2015, the TRECVID-Multimedia Event Detection (TRECVID-MED) competition evaluated multimedia event detection in videos by analyzing sounds, images, and text [1–4]. The Detection of Acoustic Scenes and Events (DCASE) competitions in 2013 [12] and 2016 [8] evaluated scene and sound recognition in audio-only recordings. The most popular standard datasets used in these investigations have allowed sound recognition to be tested in different contexts. Examples include, environmental sounds in 2015's Environmental Sound Classification - 50 (ESC-50) [13], urban sounds in 2014's Urban Sounds 8k (US8K) [14] and more recently YouTube videos in 2017's AudioSet [15] and DCASE 2017 [16]. The approaches derived from research in these datasets have shown how well we can identify acoustic content corresponding to a label.

Labels in these datasets define sounds, but rarely describe nuances of sounds. The authors of [17–19] showed that audio is associated to a subjective meaning. For instance, in the DCASE dataset [12], there were two particular labels: *quiet street* and *busy street*. Sound recognition results and confusion matrices across papers [8, 12] evidenced how although both labels defined audio from streets, the qualifier implied differences in the acoustic content. These kinds of nuances can be described

*Correspondence: bmartin1@andrew.cmu.edu
[2] Carnegie Mellon University, 5000 Forbes Ave, 15213 Pittsburgh, PA, USA
Full list of author information is available at the end of the article

with different lexical combinations. Two types that have been suggested in the literature are adjective-noun pairs (ANPs) and verb-noun pairs (VNPs).

Adjective-noun pairs can elicit a subjective meaning of the audio to the listener, defined by an adjective that shapes the natural and social environment [20]. Moreover, the subjectivity can cover other areas such as affective dimensions as explored by the authors of [18] where they collected the International Affective Digitized Sounds (IADS) dataset consisting of 111 sounds without enforcing a subjective word in the label. The sounds were presented to participants who had to categorize them into one of five classes: happiness, anger, sadness, fear, and disgust. Results showed how participants have consistent trends categorizing these sounds, suggesting a relationship between adjectives and audio content.

Verb-noun pairs can describe interactions between one or several objects and the material the objects are made of as described in [21–25]. For example, the interaction of objects and surfaces was explored in [26], where authors collected sounds corresponding to the action of drumsticks hitting and also scratching on different surfaces such as glass, wood, and metal. Results suggested acoustic differences depending on the combination of the action, defined by a verb, and the surface, defined by a noun (e.g., scratching wood and scratching metal). Moreover, authors in [19], mentioned that labeling a class using a source-action nomenclature could help reduce the expected intra-class diversity. However, this does not necessarily address the issue of inter-class diversity, which we explore in Section 4.

Investigation of adjectives and verbs as qualifiers of perceptual categories has been successfully approached in other fields. In computer vision, Borth et al. [27] introduced the VisualSentiBank to perform sentiment analysis of images [28] based on adjective-noun pairs. In video analysis, actions described by verbs have been widely explored as described in these surveys [29, 30]. In text and language processing, authors in [31] introduced SentiWordnet to perform opinion mining using adjectives. In the music domain, acoustic characteristics and lyrics have been combined to detect sentiment in [32]. It is therefore to be expected that similar exploration of audio concepts will reveal to what extent we can automatically identify such qualifying information and how it could be combined for analysis of subjectivity in multimedia content [33, 34].

In this work, we investigated for the first time the relation between audio content and both adjective-noun pair and verb-noun pair labels. The consistency between these types of pair-based labels and audio can help to analyze sentiment, affect, and opinion mining in audio as well as to complement similar pairs in other modalities such as images. Due to the lack of datasets with these types of annotations for audio, we collected, processed, and released AudioPairBank[1] under a Creative Commons (CC) license. This is a large-scale corpus consisting of 1123 pairs: 761 ANPs and 362 VNPs and over 33,000 audio files. It is based on the collaborative repository called *freesounds.org*. For this contribution, we documented the challenges and implications of collecting audio recordings with these labels. These guidelines were not previously available in the literature and now serve as a direction for researchers to further annotate or refine our annotations. We also show the degree of correlation between the audio content and the labels through sound recognition experiments and hence also providing a performance benchmark. We provide two benchmarks, binary and multi-class classification. Both yielded performance better than random, which is remarkable considering the subjective nature of our labels.

2 Collecting and processing the AudioPairBank

We start by describing the steps for collecting and processing the corpus, which is illustrated in Fig. 1. In Section 2.1, we define the list of adjective-noun and verb-noun pairs based on existing ontologies. Then, in Section 2.2, we use these labels as queries to download audio recordings from an on line repository. Finally, we refine the dataset to reduce biases, outliers, and implausible pairs in Section 2.3 to output the finalized AudioPairBank. The detailed process of each step can be seen in [35].

2.1 Selecting ANPs and VNPs based on ontologies

We looked into existing ontologies containing adjectives, nouns, and verbs to collect a list of 10,829 pairs for

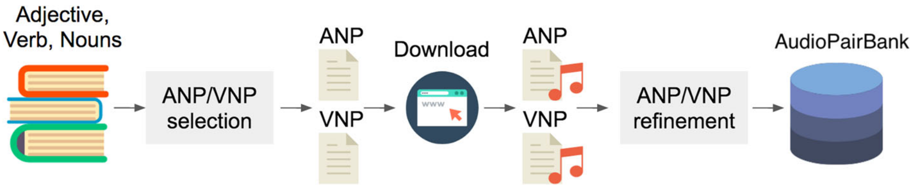

Fig. 1 Dataset construction. An overview of the collection process of AudioPairBank. The ANP and VNP labels are based on existing ontologies. The labels are used as queries to download audio from *freesound.org*. The labels and audio recordings were refined to create the final version of AudioPairbank

adjective-nouns and 9996 pairs for verb-nouns. An ontology by Davies [36] defined three audio semantic levels: sound sources, sound modifiers, and soundscape modifiers based on research where participants were asked to describe sounds using nouns, verbs, and adjectives. Additionally, Axelsson [37] suggested a list of adjectives to describe the feeling that sounds produce in individuals. Another pair of ontologies introduced by Schafer in [22] and Gygi [23] is based on soundscapes and environmental sounds, where sounds were labeled by their generating source using verbs, such as *baby crying*, and *cat meowing*. Lastly, we considered the visual sentiment ontology (VSO) presented in [27], which is a collection of ANPs based on 24 emotions defined in Plutchik's Wheel of Emotions.

After an inspection of the final list of pairs, we noticed lexical variations that were grouped. For example, comparatives and superlatives, such as *faster* and *fastest*, were grouped together as *fast*. Synonyms, such as *car*, *auto*, and *automobile*, were grouped together as *car*. Plural forms, such as *dogs* or *cars*, were grouped into their singular forms. This process implied assumptions in the audio content which may not hold true. For instance, the sound of one car is acoustically different from the sound of multiple cars. Nevertheless, grouping helped to reduce the impact of having multiple "repeated" pairs which could result in major acoustic ambiguities and low sound recognition performance.

2.2 Downloading ANPs and VNPs from the web

The list of adjective-noun and verb-noun pairs from the previous section was used to query and download audio recordings from *freesound.org*. The website has the largest audio-only archive of sounds with around 230,000 recordings, which allowed us to collect audio in a large-scale. Moreover, the website has been successfully employed before to create popular datasets (ESC-50, Freefield, US8K) [13, 14, 38]. Other websites such as *soundcloud.com* and *findsounds.com* were considered, but not employed because they either contained mainly music or had less sound recordings in their archives.

The *freesound.org* website is a collaborative repository of audio where users upload recordings and write tags to describe their content. A tag is a keyword that describes and highlights the content of the audio recording [39]. This folksonomy-like structure of repository has the benefit of reflecting the popular and long-tailed combinations of tags. In this manner, we can observe what are the socially-relevant adjectives, verbs, and nouns.

The tags of a given recording are combined to create weak labels of adjective-noun pairs and verb-noun pairs. The weak labels happen because users upload audio recordings and provide tags based on what they consider relevant. However, tags do not follow a particular order, may describe only a portion of the content and are not accompanied with the time-stamp of the sound-tag occurrence. Moreover, the order of the tags could influence the meaning, such as tag with a verb intended to be used as an adjective. The intent of the user is unknown, but its exploration is necessary for large-scale collection and analysis. We also expect machine learning algorithms to help us determine the degree of the relation between such pairs and their corresponding sounds.

2.3 Refining the downloaded ANPs and VNPs

The downloaded audio recordings along with their labels revealed several characteristics discussed in this section. Therefore, we refined the corpus with the goal of increasing the quality and diversity of audio concept pairs. The process is illustrated in Fig. 2. A manual revision has been employed by other authors [15, 27] to improve their automatically collected datasets.

Minimum sampling rate: The chosen rate was 16 kHz because 95% of the files were 16 kHz and 44.1 kHz. Files with a lower rate were discarded because such rates reduce the frequency information in the recording.

Contain special words: We removed pairs tagged with words that implied unwanted content. For example, *loop*,

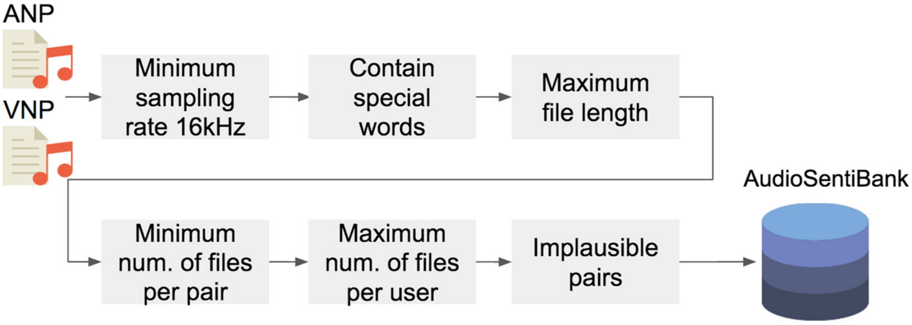
ANP
VNP
Minimum sampling rate 16kHz
Contain special words
Maximum file length
AudioSentiBank
Minimum num. of files per pair
Maximum num. of files per user
Implausible pairs

Fig. 2 Data filtering. The refining process consisted of six steps where pairs that did not meet a certain criterion were discarded. Steps 2,5 and 6 were responsible of the highest number of discarded pairs

loops, or *looping* contained sounds which repeated over and over again. The repetition or periodicity was artificial and could mislead the sound recognition systems. We also removed pairs with words such as *sound*, *audio*, or *effect* because they did not add meaning. Pairs with the word *processed* were removed because the audio files commonly contained music or overlapped music throughout the recording. We also removed pairs with redundancy of terms such as *noisy noise* or *natural nature*. Another discarded pattern happened with terms related to music genre such as *heavy metal* and *classic rap* or types such as *waving techno* and *ringing music*. Nevertheless, we kept some pairs related to music such as *happy music*, *sad music*, or *dramatic guitar*. Another pattern happened with sound packs, which consisted of groups of audio files, with exactly the same tags and uploaded by the same user. However, not all audio content from every file was related to the tags. Because these bundles did not occur often and are hard to track automatically, we removed the obvious ones, but perhaps kept others.

Maximum audio file length: We removed recordings with long duration because longer audio files had audio content that was not described by the tags. We computed the distribution of the duration for each pair and removed the outliers in the distribution based on Tukey's range test[2]. The outliers are values larger than the third quartile (Q3) of the distribution plus 1.5 times the interquartile range (IQR), and formally: outlier $> Q3 + 1.5 \times IQR$. For ANPs, outliers have durations greater than 129.5 s ($54.5 + (1.5 \times 50)$), and for VNPs greater than 67.5 s ($28.5 + (1.5 \times 26)$). About 80% of the files are up to 20 s in length for ANPs and VNPs.

Minimum number of files per pair: We discarded pairs with less than 20 files. Such low count was an indicator of rare pairs, which might not be worth exploring in this work. The minimum number is consistent with other available datasets [8, 40].

Maximum number of audio files per user: For some pairs, there were users who dominated the contribution of audio recordings. Users tend to use similar recording devices and conditions which can cause a data bias. As a consequence, machine learning-based algorithms for sound recognition can learn these biases instead of the audio content as demonstrated in [41]. To reduce such user-specific influence, we allowed a maximum of 25% of recordings per pair to be from any single user.

Implausible pairs: manual inspection and plausibility score:

We employed manual inspection to catch salient implausible pairs. We complemented this approach with a data-driven metric to determine the degree of plausibility. Both approaches were responsible of discarding 12% of the ANPs 51% of VNPs as shown in Fig. 3 and some examples are in Table 1.

We manually inspected the audio concept pairs and discarded those which were implausible and could not be associated to an acoustic semantic meaning, for example, some ANPs derived from Plutchik's Wheel of Emotion, such as *slow fear*, a sound that is arguably impossible to reproduce. This problem was also faced in the visual sentiment ontology [27], and pairs were discarded, such as

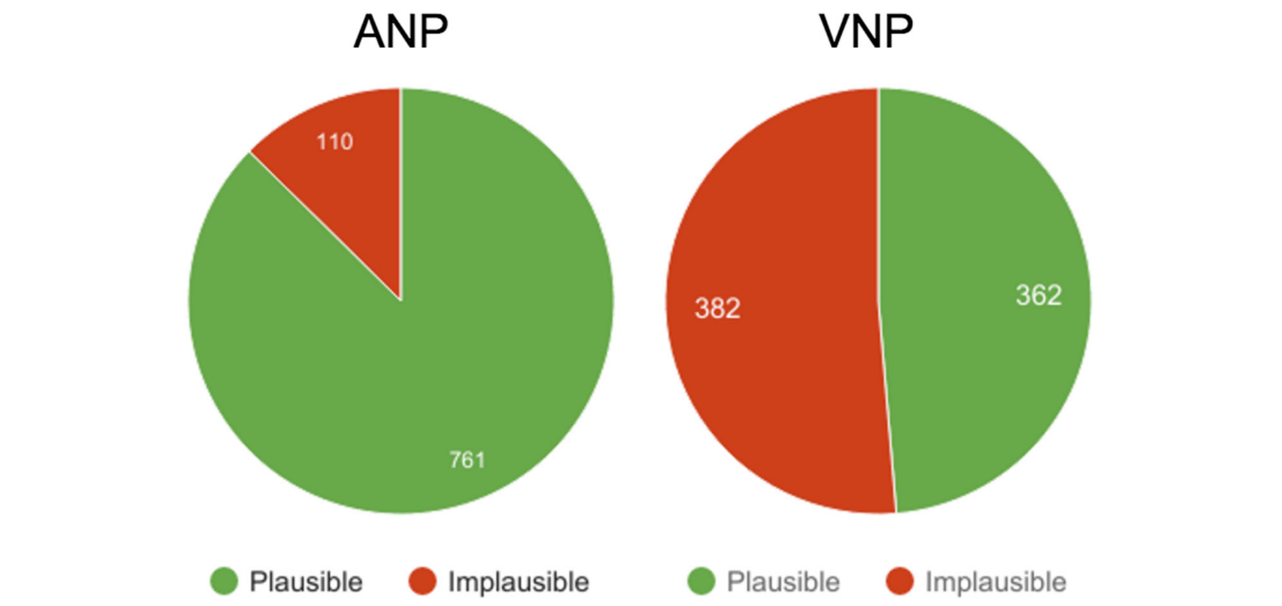

Fig. 3 Plausible and implausible pairs. The charts show the number of plausible and implausible pairs after using both methods, manual inspection and plausibility score. In contrast to adjectives, verbs have more restrictions in how they can be combined with nouns to convey plausible meaning

Table 1 Examples of plausible and implausible pairs

Plausible	Implausible
ANPs	
Happy music	Windy bird
Slow car	Industrial hands
Echoing footsteps	Slow fear
Echoing alert	Extreme noise
VNPs	
Singing bird	Laughing animation
Crying baby	Falling autum
Flying bee	Clapping hat
Honking car	Talking text

Note how for example "fast food" is an implausible pair because despite that it could be associated to a semantic meaning, it does not convey an acoustic semantic meaning

fresh food or *favorite book*. Other examples of implausible pairs are *walking winter* and *singing park*, where nouns define a time and place, but were implausible because of the verb. Some implausible nouns like *future* or *design* represented abstract meanings and were hardly connected to consistent sounds. An interesting pair also discarded was *talking bird*, which appeared to be semantically wrong, but possible. A closer look into the corresponding audio files revealed recordings of a bird talking, a parrot perhaps, but these recordings were rare and the majority of the recordings contained talking people with bird sounds in the background.

The plausibility score (PS) was designed to favor diversity of users, number of files, and uniqueness of files for the given pair. For example, the manually discarded pair *singing park* was selected because a park cannot sing. This pair might have emerged because files were tagged with different words such as *singing*, *walking*, *relaxing*, *park*, *bird*, and *people*. Although the pair *singing park* occurred in several audio files, it never occurred together in any other recording and was rather always together with other pairs such as *singing bird* and *walking people* and *relaxing park*. Hence, these kind of pairs yielded a low plausibility score. The score is defined as follows:

$$PS(cp) = \frac{\frac{u_{cp}}{n_{cp}} + \frac{f_{cp}}{n_{cp}}}{2} = \frac{u_{cp} + f_{cp}}{2n_{cp}} \qquad (1)$$

Here, n_{cp} is the total number of files belonging to a concept pair *cp*. Then, u_{cp} is the number of unique users that uploaded files for a concept pair *cp* and f_{cp} is the number of files that are unique to a concept pair *cp*. A file is unique to a *cp* if it is not tagged with any other concept pair in the existing set. The division by 2 is necessary because for a perfectly plausible concept pair the numerator becomes $1+1$, and so the division by two will keep the metric in the range between 0 (least plausible) and 1 (most plausible). We observed that rare pairs such as the ones described in the previous paragraphs obtained a score lower than 0.2.

2.4 Finalized AudioPairBank

The refined corpus is one of the largest available datasets for sounds and the only dataset with adjective-noun and verb-noun labels. The ANPs and VNPs are weak labels based on the collaborative repository in *freesound.org*. The main statistics of the corpus are included in Table 2. AudioPairBank consists of 761 ANPs and 362 VNPs for a total of 1123 pairs. One or more pairs can correspond to the same audio file. The number of shared-unique audio files are 58,626–16,335 for ANPs and 38,174–20,279 for VNPs for a total of 96,800–33,241 files. The average number of unique files are 21 for ANPs and 56 for VNPs. The number of unique nouns is 1187, unique verbs is 39, and unique adjectives is 75. The influence of a user per pair was set to a maximum of 25% contribution of files. Note that the total number of unique files and users does not correspond to the sum of the previous rows because some files are repeated in both categories. Regarding the length, ANPs had a larger file duration than VNPs with almost twice the length. The last column shows the total disk space of the audio files in waveform audio file (WAV) format.

3 Analysis of AudioPairBank

We performed an analysis on different aspects of ANPs and VNPs, such as co-occurrences of adjective, verb, nouns, duration of audio recordings, number of audio files, number of tags, and number of users.

3.1 Number of audio files per ANP and VNP

The distribution of the number of files per pair for ANPs and VNPs gives us an intuition of which pairs are more common for users on the web. Figure 4 shows a decreasing distribution with a long tail trimmed due to space limitations. ANPs show a smoother distribution decrement, which translates to a more uniform number of files per concept pair.

3.2 Number of users per ANP and VNP

We looked at how many different users contributed to the pairs to have an intuition about the diversity of users

Table 2 ANPs double the number of VNPs

	Pairs	Total - Unique files	Users	Hours	Size
ANPs	761	58,626 - 16,335	2,540	892	528
VNPs	362	38,174 - 20,279	3,279	375	212
Total	1123	96,800 - 33,241	4,478	1,267	740

Audio content is generally described with multiple adjectives. Whereas verbs are used to describing actions, which happen with less frequency and typically refer to the source of the dominant sound. Size is counted in gigabyte (GB)

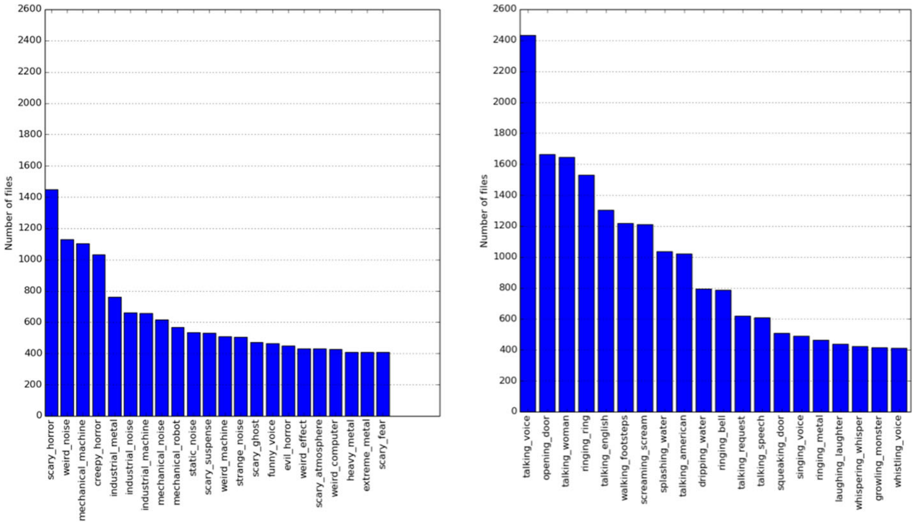

Fig. 4 Number of files per pair. ANPs (left) and VNPs (right) with the largest number of files. Both plots show a decreasing distribution with a long tail trimmed due to space limitations. ANPs show a smoother decrement, which translates to a more uniform number of files per concept pair

per concept pair in Fig. 5. The figure gives a first intuition of which concept pairs are more diverse, but it does not show if pairs are equally distributed among contributors. Hence, Fig. 6 helps to visualize how the files per pair are distributed among users. Most concept pairs are very diverse with respect to the uploading users, but a few are more dominated by individual users.

We also observed that most users commonly employed a small set of tags with high frequency. The most frequent tags across users are shown in Table 3. Rarely occurring tags often come from single users.

3.3 Duration of ANP and VNP audio files

Adjectives, in some cases, suggest the duration of the recording. For example on average, audio containing *calm*, *rural*, *peaceful*, and *quiet* had longer durations (more than 5 min) than those tagged with *accelerating* or *rushing* (less than 2 min).

Verbs describe actions, and hence, we expected the duration of the audio recordings of VNPs to correspond to the approximate length of the described action. However, even if most actions lasted between 1 and 5 s, the median duration of VNP audio files was around 10 s. This may

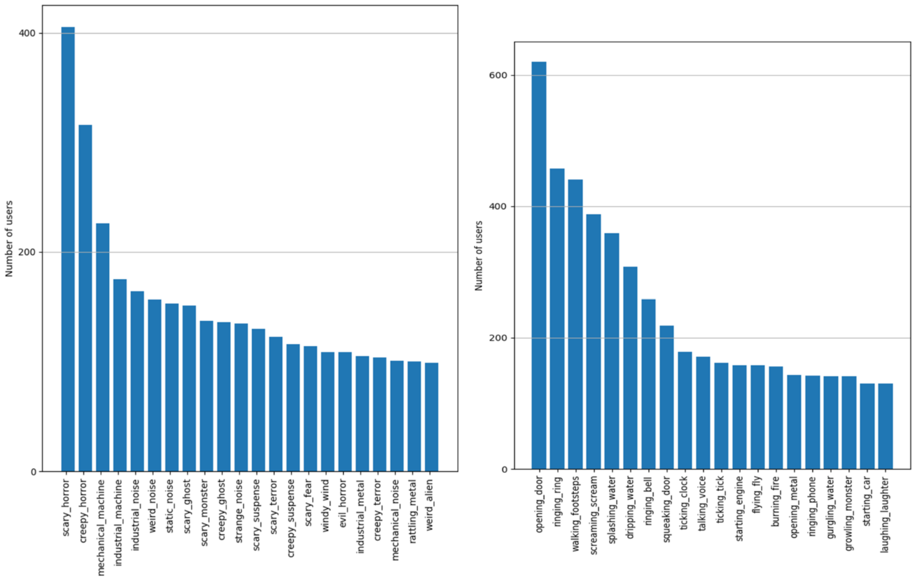

Fig. 5 Number of users per pair. Number of users that uploaded files for ANPs (left) and VNPs (right). We show examples of pairs with at least 400 files

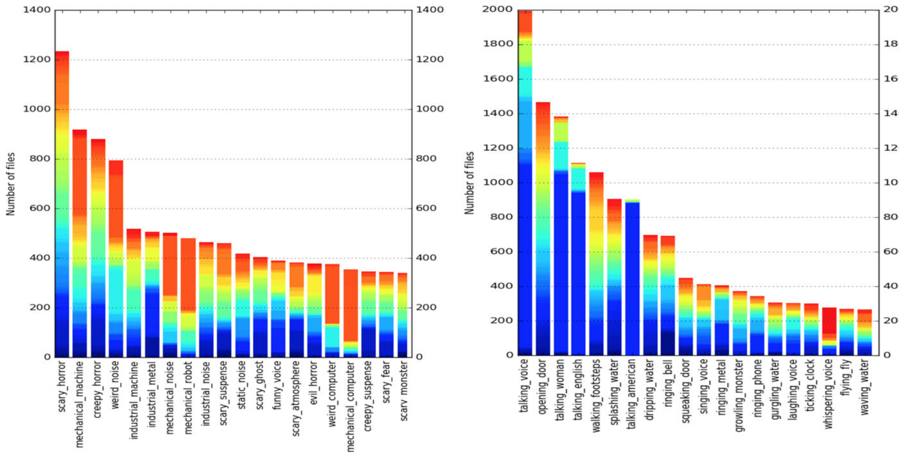

Fig. 6 Number of files per concept pair color-coded by the uploading users. ANPs (left) and VNPs (right) with the highest number of contributors (up to 400). Each user that contributed to a concept pair is color-coded. Most concept pairs are diverse, but a few may have dominant users

happen because the action described by the verb tend to occur more than once, or because other actions take place within the recording.

Nouns have more length variability that depends on what it is describing. Locations, environments, and field recordings are associated to longer durations, such as *city*, *market*, *beach*, or *rain*, while objects are associated to shorter durations, such as *cup* or *door*.

3.4 Correlation between the number of tags and the length of the ANP and VNP audio files

We expected that more tags will describe more audio content and therefore, correlate with longer durations. However, we found a weak to almost no correlation between both, as illustrated in Fig. 7. We validated our observations by computing Spearman's rank correlation coefficient (SRCC), which is a non-parametric measure of statistical dependence between two variables. An SRCC value close to zero in combination with a very small p value indicates no correlation in the data. In our case, the SRCC value for ANPs is 0.073 with a p value of 5.358e−97, and for VNPs is 0.0149 with a p value of 6.31e−05. For each file, the average number of tags is 15. Nevertheless, the number of tags together with the duration of the audio file suggested a distinction between acoustic scenes/soundscapes and sounds.

Table 3 High-frequency tags common across users

	Tags
Adjectives	Scary, creepy, industrial, funny
Verbs	Talking, walking, laughing, singing,
Nouns	Atmosphere, horror, noise, voice

3.5 Co-occurrences of tags in ANP and VNP audio files

In order to understand the context in which concept pairs occur, we analyzed the co-occurrences of the accompanying adjectives, nouns, and verbs tags within the audio file.

Some adjectives occur more frequently than others, such as *loud*, *heavy*, *scary*, and *noisy* in contrast to *exotic*. Adjectives that occur frequently tend to describe nouns that are commonly locations: *landscape*, *coast*, or *nature*. As expected, we found almost no tags using colors as adjectives. An interesting co-occurrence of adjectives was when they had opposite meaning such as *slow* and *fast* or *peaceful* and *loud*. After manual inspection of audio files, we concluded that this was an indicator of changes in the audio content throughout the recording, specially when the noun was shared. For example, a recording had two pairs, *slow train* and *fast train*, which had a train passing slowly and then followed by another one passing at high-speed.

Verbs tend to occur less frequently than adjectives and with a more restricted set of nouns. For instance, *flying* occurs mainly with *airplane*, *engine*, *bird*, and *helicopter*. In a similar manner, almost the same verbs occur with human-related nouns such as *baby*, *child*, *man*, and *woman*. One pattern observed for verbs is that they may describe and complement another verb. For example, *open* and *close* co-occur with *squeaking* and *banging*, all with the noun door. These combinations specify how the door was opened or closed. Similar to adjectives-adjectives, verbs could be an indicator of changes in the audio content throughout the recording, such as *singing* and *clapping* described a music concert.

Nouns were more helpful to provide context and clarify the sound source. An example of context are *thunder*, *rain*, and *wind*, which described the acoustics of a storm,

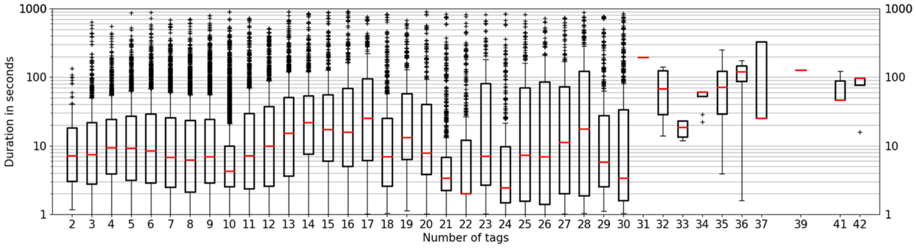

Fig. 7 Number of tags vs duration of files corresponding to ANPs with up to 43 tags. We expected that more tags will describe more audio content and therefore, correlate with longer durations. However, the correlation between number of tags and duration was weak or none. For instance, files with 16 and 27 tags will have the same average duration. A similar trend was observed for VNPs

also *frogs* and *insects* and *water* described the acoustics of a Savannah. Moreover, users sometimes included the location such as *Florida* or the time such as *day* and *night* or the season such as *spring*. An example of sound source is when the sound of an *engine* happened with *car* or *train* or *airplane* or *ship*, and thus, we knew the specific source of the engine sound. Importantly, the noun *noise*, which co-occurs very frequently with other tags, sometimes indicates an unintelligible sound, but was more commonly employed to define sounds happening in the background which were unrelated to the target sound.

4 Experimental setup for benchmark computation of AudioPairBank

The previous sections described the challenges of collecting acoustic pairs, such as weak labeling from folksonomies. In addition, we also refined our pairs expecting to find consistency between the acoustics and the labels describing the actions (verbs) and properties (adjectives) of the sounds. Therefore, in this section, we describe the setup of the two main experiments on the audio of the adjective-noun pairs and verb-noun pairs contained in AudioPairbank. First, the input audio was standardized into the same format and passed through a feature extraction step. Next, the extracted features were passed through a system for binary classification and multi-class classification.

The audio files from the dataset needed to be standardized into the same format. We chose WAV format, sampling rate of 44.1 kHz, encoding pulse code modulation (PCM) 16 bits and one channel. These parameters were already dominant in the corpus and are also common in the available datasets. Moreover, audio files have variable length, and thus, each file was trimmed into 4 s segments with a 50% overlap. Files with shorter duration were adjusted during the feature computation. These two parameters yielded the best sound event classification results among different values [14, 42]. The AudioPairbank corpus has three partitions, training, cross validation (CV), and testing with a ratio of 40%-30%-30%.

For each audio of 4 s, we extracted mel frequency cepstral coefficients (MFCCs) features because they provide a competitive baseline performance. We used the toolbox Yaafe [43] to compute MFCCs with 13 and 20 coefficients and appended their first and second derivatives, also called delta and double delta. The window size was 30 ms at every 10 ms. Each MFCC frame including the deltas and double deltas were stacked into a single vector. Stacking is used to consider the temporal context [44]. Files with duration shorter than 4 s were padded with frames containing zeros.

Binary classification allows the classifier to decide whether a test pair-class belongs to the trained class or not. This is useful when dealing with negative samples and unlabeled data as in the TRECVID-MED Evaluations [45]. It is also useful in terms of computational efficiency—it is faster to retrain binary models (hours) in contrast to multi-class models (weeks), and easier to parallelize. We used a one-vs-all setup using a support vector machines (SVM) [46] classifier. We trained one SVM for each ANP and VNP using 100 segments corresponding to the positive (target) pair-class and 200 corresponding to a negative (not-the-target) pair-class. The 100 segments were randomly selected from the pool of the positive pair-class. This number corresponds to the minimum number of segments per pair-class. The 200 segments were randomly selected from the pool of other classes avoiding repeated files. The SVM had a linear kernel, and using CV set, we tuned the soft margin parameter *C* with the following values: 5.0, 2.0, 1.0, 0.5, and 0.01, with five yielding the best results.

Multi-class classification has to discern between multiple pair-classes and forces every input audio to belong to one of the trained pair-classes. This classification type is more complicated than the previous given the large number and ambiguity of the classes. We employed two different algorithms. First, we used a multi-class random forest (RF) [46] classifier used to compute baseline performance in [9, 40]. RF are an ensemble learning method that operates by creating multiple decision trees at training time.

Then, at testing time, each tree votes to predict a class, minimizing over fitting. We trained two RF: one for all the VNPs and one for all the ANPs using 100 segments corresponding to each class. The 100 segments were randomly selected from the pool of positive segments. The two RF were tuned using CV set to find the number of trees, we tried 5, 10, 20, 50, and 100, where 100 yielded the best results. Using more than 100 trees surpassed the capabilities of our computing resources. Second, we trained a convolutional neural networks (CNN), which forms the basis of the state of the art in sound classification. In [40], the log-mel spectrogram of each sound event recording is treated as an image and passed to the CNNs. We employed the same architecture. The first convolutional rectifier linear unit (ReLU) layer consisted of 80 filters of rectangular shape (57 × 6 size, 1 × 1 stride) allowing for slight frequency invariance. Max pooling was applied with a pool shape of 4 × 3 an stride of 1 × 3. A second convolutional ReLU layer consisted of 80 filters (1 × 3 size, 1 × 1 stride) with max pooling (1 × 3 pool size, 1 × 3 pool slide). Further processing was applied through two fully connected hidden layers of 5000 neurons with ReLU non-linearity. The final output is a softmax layer. Training was performed using Keras implementation of mini-batch stochastic gradient descent with shuffled sequential batches (batch size 1000) and a Nesterov's momentum of 0.9. We used L2 weight decay of 0.001 for each layer and dropout probability of 0.5 for all layers.

To evaluate binary classification performance we computed accuracy (Acc), f-score, and area under the curve (AUC). For multi-class classification, we computed accuracy. For all the metrics, we computed micro-averaging with the toolbox sci-kit learn [46].

5 Results and discussion

The overall binary classification performance of ANPs and VNPs was better than expected. As a reference, similar experiments with AudioSet [47] had comparable performance (AUC 85%). AudioSet has 485 sound classes and did not deal with the subjectivity treated in this paper. The performance for both types of pairs is shown in Table 4 with the following results: accuracy of 69 and 71%, f-score of 51 and 53%, and AUC of 70 and 72%. The results correspond to the extracted audio features of 13 MFCCs;, expanding this to 20 MFCCs did not provide a performance gain. One explanation why VNPs performed better is because verbs tend to be less subjective or more neutral than adjectives [48]. This means that the acoustic characteristics may be more distinguishable for classifiers as it is for humans. For example, people might argue about what would be the sound of a *beautiful car*, but not so much about the sound of a *passing car*.

Table 4 Overall binary classification performance

	Features	Acc%	f-score%	AUC%
ANP	13 MFCCs+Δ+ΔΔ	69	51	70
VNP	13 MFCCs+Δ+ΔΔ	71	53	72

We show the best and worst performing ANPs and VNPs in Table 5. The best detected ANPs had over 93% accuracy and corresponded to distinguishable sounds of phone numbers being pressed or phones' tones, such as *industrial phone*, *echoing phone*, and *weird cell*. On the contrary, *noisy glitch* and *extreme noise* had accuracy around 1% and corresponded to pairs with adjectives that described a generic meaning rather than specific. The best detected VNPs were *howling dog*, *howling wolf*, *crying insects*, and *howling animal* with accuracy greater than 94%. These sounds tend to have almost no overlapping audio. On the contrary, *splashing water*, *crackling footsteps*, *splashing river*, *gurgling water*, and *breaking snow* had accuracy around 2%. These wide-band, background-noise-like continuous sounds are hard to classify. A similar problematic arose in [14] with sounds such as *air conditioning* and *engine idling*. Additionally, pairs corresponding to long duration recordings, commonly related to environmental sounds and field recordings, tend to have lower performances.

We looked at the overall performance of pairs sharing a common adjective, verb or noun to estimate how well we could detect them. The adjective with the highest accuracy, 71%, corresponded to *industrial*, which commonly paired with nouns such as *hands*, *phone*, and *metal*. On the other hand, the verb with the highest accuracy, 73%, corresponded to *singing*, which commonly paired with *choir*, *crowd*, *man*, *child*, and *woman*. Nouns such as *alert*, *phone*, and *guitar* performed well for different adjectives and verbs because they have a specific timbre.

Table 5 Top five best and worst binary classified ANPs and VNPs

Best	Worst
ANPs	
Weird cup	Funny english
Industrial phone	Heavy rain
Echoing phone	Noisy glitch
Echoing alert	Extreme noise
Weird cell	Loud fireworks
VNPs	
Howling dog	Splashing water
Howling wolf	Crackling footsteps
Crying insects	Splashing river
Howling animal	Gurgling water
Ringing cup	Breaking snow

The overall multi-class classification of ANPs and VNPs yielded good performance. The accuracies shown in Table 6 are respectively for RF and CNN, 1.6 and 2.1% for ANPs and 4.5 and 7.4% for VNPs. While the multi-class performance is lower than the detection experiments, it is still higher than random performance, which corresponds to 0.13% for ANPs and 0.27% for VNPs. As supported in the literature [40], the CNN outperformed the RF algorithm. Similar to the detection case, the RF numbers correspond to audio features of 13 MFCCs. The performance difference between ANPs and VNPs may be explained because there are more than twice number of ANPs than there are VNPs, which makes the classification task harder for ANPs because the classifier has to discern between more pair-classes.

An issue with the multi-class setup is that some audio files corresponded to more than one pair label. When a classifier is trained sharing one or more audio files for one or more classes, it struggles to define a decision boundary to separate the classes. A solution could be to add more acoustically diverse training audio files, which could ameliorate the problem by aiding the classifier to generalize the class boundaries. Nevertheless, this is an issue that has to be further explored because it is expected that sounds can be labeled with more than one adjective, verb or noun.

The multi-class setup also allowed us to observe pairs that were commonly confused by their acoustic characteristics as shown in Table 7. The confusion matrix is not included here for lack of space, but in general, the confusions look conceptually reasonable for both types of pairs. For the ANPs, confusions happened when ANPs shared the same adjective and when pairs shared similar contexts expressed by the noun, such as in *extreme rain* and *heavy thunder*. A similar case was observed with VNPs, where the confusions happened when the verb and the noun expressed similar meanings, such as*splashing lake* and *walking river*.

These experiments and results evidence a degree of consistency between tag-pairs and the sounds associated to them, despite of the setup limitations and assumptions. In this work, we take a first step towards exploiting tag-pair-based actions and properties of sounds automatically from a large-scale repository. We point out some of the challenges, which have not been published to the best of our knowledge, and could be of great use for research if we want to take advantage of the massive amounts of web audio and video recordings. Other lexical combinations could be explored, such as other verb conjugations and adverbs to add meaning to a given action. Also, sequences of lexical combinations may help describe the order actions take place in a specific scene or describe the properties of the acoustic scene. Reliable recognition of the acoustic nuances can benefit several applications. For example, VNPs can support violent detection in image processing [49]. In another example, ANPs could be used for opinion mining to determine the quality of urban soundscapes [50]. Similarly, ANPs could also be combined with the image-based ANPs [28] for sentiment analysis of videos. Hence, we encourage further exploration of the nuances, both of audio-only recordings and as a complement of similar research on images and text in multimedia analysis.

Table 6 Overall multi-class classification performance for random forest and CNN

Classifier	Pair	Features	Accuracy%
RF	ANP	13 MFCCs+Δ+ΔΔ	1.6
RF	VNP	13 MFCCs+Δ+ΔΔ	4.5
CNN	ANP	13 MFCCs+Δ+ΔΔ	2.1
CNN	VNP	13 MFCCs+Δ+ΔΔ	7.4

Table 7 Each row correspond to an example of pairs that were highly confused

ANPs	
Extreme rain	Heavy thunder
Heavy thunder	Heavy wind
Distant rain	Distant thunder
Relaxing water	Relaxing creek
Echoing church bell	Echoing hall
VNPs	
Passing railway	Passing train
Singing bird	Tweeting bird
Talking crowd	Walking noise
Splashing lake	Waving river
Burning fire	Crackling fire

6 Conclusions

In previous years, sound recognition has been used to identify sounds also called audio concepts. However, sounds have nuances that may be better described by adjective-noun pairs such as *breaking glass*, and verb-noun pairs such as *calm waves*. Datasets with these types of labels are unavailable. In this work, we provide an investigation about the relation between audio content and weak labels corresponding to adjective-noun pairs and verb-noun pairs. For this study, we collected, processed, and made available, AudioPairBank, a large-scale corpus consisting of 761 ANPs and 362 VNPs corresponding to over 33,000 audio files. Using this dataset, we evaluated classification performance of audio recordings and the results supported a degree of consistency between tag-pairs and sounds. We provided a benchmark better than random performance, regardless of complications such as using weak labels and the tagging assumptions from a collaborative repository. We expect to guide researchers exploring labels and sounds. We showed an initial performance for classifying these types of lexical pairs and encourage the exploration of other lexical combinations. Moreover, we expect to further research that analyzes sentiment, affect, and opinion mining in audio and multimedia recordings.

Endnotes

[1] http://audiopairbank.dfki.de/

[2] http://en.wikipedia.org/wiki/Tukey%27s_range_test

Abbreviations

Acc: Accuracy; ANP: Adjective-noun pair; ANPs: Adjective-noun pairs; AUC: Area under the curve; CC: Creative commons; CNN: Convolutional neural networks; CV: Cross validation; DCASE: Detection of acoustic scenes and events; ESC-50: Environmental sound classification - 50; GB: Gigabyte; IADS: International Affective Digitized Sounds; IQR: Interquartile range; MFCCs: Mel frequency cepstral coefficients; PCM: Pulse code modulation; PS: Plausibility score; Q3: Third quartile; ReLU: Rectifier linear unit; RF: Random forest; SRCC: Spearman's rank correlation coefficient; SVM: Support vector machines; TRECVID-MED: TRECVID-multimedia event detection; US8K: Urban sounds 8k; VNP: Verb-noun pair; VNPs: Verb-noun pairs; VSO: Visual sentiment ontology; WAV: Waveform audio file

Availability of data and materials

http://audiopairbank.dfki.de

Authors' contributions

SS conducted the collection of the data and the experimentation. BE wrote the article and provided the audio-related knowledge needed for the data collection, experimentation, and conclusions. DB supervised Sebastian as his master student and provided his expertise on a similar research applied to computer vision. BR and IL were BE's advisors and provided supervision and funding. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]University of Kaiserslautern, DFKI, Kaiserslautern, Germany. [2]Carnegie Mellon University, 5000 Forbes Ave, 15213 Pittsburgh, PA, USA.

Received: 24 January 2018 Accepted: 26 August 2018
Published online: 15 September 2018

References

1. P. Schäuble, *Multimedia information retrieval: content-based information retrieval from large text and audio databases*, vol. 397. (Springer Science & Business Media, 2012)
2. P. Natarajan, P. Natarajan, S. Wu, X. Zhuang, A. Vazquez Reina, S. N. Vitaladevuni, K. Tsourides, C. Andersen, R. Prasad, G. Ye, D. Liu, S.-F. Chang, I. Saleemi, M. Shah, Y. Ng, B. White, L. Davis, A. Gupta, I. Haritaoglu, in *Proceedings of TRECVID 2012*. BBN VISER TRECVID 2012 multimedia event detection and multimedia event recounting systems (NIST, USA, 2012)
3. Z. Lan, L. Jiang, S.-I. Yu, C. Gao, S. Rawat, Y. Cai, S. Xu, H. Shen, X. Li, Y. Wang, W. Sze, Y. Yan, Z. Ma, N. Ballas, D. Meng, W. Tong, Y. Yang, S. Burger, F. Metze, R. Singh, B. Raj, R. Stern, T. Mitamura, E. Nyberg, A. Hauptmann, in *Proceedings of TRECVID 2013*. Informedia @ TRECVID 2013 (NIST, USA, 2013)
4. H. Cheng, J. Liu, S. Ali, O. Javed, Q. Yu, A. Tamrakar, A. Divakaran, H. S. Sawhney, R. Manmatha, J. Allan, et al., in *Proceedings of TRECVID*. Sri-sarnoff aurora system at trecvid 2012: Multimedia event detection and recounting, (2012)
5. J. Maxime, X. Alameda-Pineda, L. Girin, R. Horaud, in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. Sound representation and classification benchmark for domestic robots (IEEE, 2014), pp. 6285–6292
6. M. Janvier, X. Alameda-Pineda, L. Girinz, R. Horaud, in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*. Sound-event recognition with a companion humanoid (IEEE, 2012), pp. 104–111
7. W. K. Edwards, E. D. Mynatt, in *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology, UIST '94*. An architecture for transforming graphical interfaces (ACM, New York, 1994), pp. 39–47. https://doi.org/10.1145/192426.192443
8. A. Mesaros, T. Heittola, T. Virtanen, in *Signal Processing Conference (EUSIPCO), 2016 24th European*. TUT database for acoustic scene classification and sound event detection (IEEE, Budapest, 2016), pp. 1128–1132
9. J. Salamon, C. Jacoby, J. P. Bello, in *Proceedings of the 22nd ACM international conference on Multimedia*. A dataset and taxonomy for urban sound research, (Orlando, 2014), pp. 1041–1044
10. M. Yang, J. Kang, Psychoacoustical evaluation of natural and urban sounds in soundscapes. J. Acoust. Soc. Am. **134**(1), 840–851 (2013)
11. K. Hiramatsu, K. Minoura, in *Proc. Internoise*. Response to urban sounds in relation to the residents' connection with the sound sources (Societé Française d'Acoustique. CD-Rom. Niza (Francia), 2000)
12. D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, M. D. Plumbley, *Detection and classification of acoustic scenes and events: an IEEE AASP challenge*
13. K. J. Piczak, in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30*. ESC: dataset for environmental sound classification, (2015)
14. J. Salamon, C. Jacoby, J. P. Bello, in *Proceedings of the 22nd ACM international conference on Multimedia*. A dataset and taxonomy for urban sound research, (Orlando, 2014), pp. 1041–1044
15. J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter, in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. Audio set: an ontology and human-labeled dataset for audio events (IEEE, New Orleans, 2017), pp. 776–780
16. A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, T. Virtanen, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*. DCASE 2017 challenge setup: tasks, datasets and baseline system, (2017)
17. S. Ntalampiras, A transfer learning framework for predicting the emotional content of generalized sound events. J. Acoust. Soc. Am. **141**(3), 1694–1701 (2017)
18. R. A. Stevenson, T. W. James, Affective auditory stimuli: characterization of the International Affective Digitized Sounds (IADS) by discrete emotional categories. Behav. Res. Methods. **40**(1), 315–321 (2008)
19. G. Lafay, M. Lagrange, M. Rossignol, E. Benetos, A. Roebel, A morphological model for simulating acoustic scenes and its application to sound event detection. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(10), 1854–1864 (2016)
20. S. Frühholz, W. Trost, S. A. Kotz, The sound of emotions—towards a unifying neural network perspective of affective sound processing. Neurosci. Biobehav. Rev. **68**, 96–110 (2016)
21. A. Darvishi, E. Munteanu, V. Guggiana, H. Schauer, M. Motavalli, M. Rauterberg, in *Human–Computer Interaction*. Designing environmental sounds based on the results of interaction between objects in the real world (Springer, Boston, 1995), pp. 38–42
22. R. M. Schafer, *The soundscape: our sonic environment and the tuning of the world*. (Inner Traditions/Bear, 1993)
23. B. Gygi, G. R. Kidd, C. S. Watson, Spectral-temporal factors in the identification of environmental sounds. J. Acoust. Soc. Am. **115**(3), 1252–1265 (2004)
24. J. A. Ballas, J. H. Howard Jr, Interpreting the language of environmental sounds. Environ. Behav. **19**(1), 91–114 (1987)
25. D. Dubois, C. Guastavino, M. Raimbault, A cognitive approach to urban soundscapes: using verbal data to access everyday life auditory categories. Acta Acustica U. Acustica. **92**(6), 865–874 (2006)

26. A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, W. T. Freeman, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Visually indicated sounds, (2016)
27. D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, in *Proceedings of the 21st ACM International Conference on Multimedia*. Large-scale visual sentiment ontology and detectors using adjective noun pairs, MM '13 (ACM, New York, 2013), pp. 223–232
28. T. Chen, D. Borth, T. Darrell, S.-F. Chang, Deepsentibank: visual sentiment concept classification with deep convolutional neural networks. arXiv preprint arXiv:1410.8586 (2014)
29. J. M. Chaquet, E. J. Carmona, A. Fernández-Caballero, A survey of video datasets for human action and activity recognition. Comp. Vision Image Underst. **117**(6), 633–659 (2013)
30. R. Poppe, A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
31. S. Baccianella, A. Esuli, F. Sebastiani, in *Lrec*. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, vol. 10, (2010), pp. 2200–2204
32. J. Zhong, Y. Cheng, S. Yang, L. Wen, Music sentiment classification integrating audio with lyrics. J. Inf. Comput. Sci. **9**(1), 35–44 (2012)
33. R. W. Picard, Computer learning of subjectivity. ACM Comput. Surv. (CSUR). **27**(4), 621–623 (1995)
34. M. Soleymani, Y.-H. Yang, Y.-G. Jiang, S.-F. Chang, in *Proceedings of the 23rd ACM International Conference on Multimedia*. Asm'15: The 1st international workshop on affect and sentiment in multimedia (ACM, New York, 2015), pp. 1349–1349
35. S. Sager, *Audiopairbank - large-scale vocabulary for audio concepts and detectors, Master's thesis*. (Technische Universitat KaisersLautern, 2016)
36. W. Davies, M. Adams, N. Bruce, R. Cain, A. Carlyle, P. Cusack, D. Hall, K. Hume, A. Irwin, P. Jennings, M. Marselle, C. Plack, J. Poxon, Perception of soundscapes: An interdisciplinary approach. Appl. Acoust. **74**(2), 224–231 (2013)
37. Ö. Axelsson, M. E. Nilsson, B. Berglund, A principal components model of soundscape perception. J. Acoust. Soc. Am. **128**(5), 2836–2846 (2010)
38. D. Stowell, M. Plumbley, in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. An open dataset for research on audio field recording archives: freefield1010, (2014)
39. B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L.-J. Li, Yfcc100m: The new data in multimedia research. Commun. ACM. **59**(2), 64–73 (2016)
40. K. J. Piczak, in *IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Environmental sound classification with convolutional neural networks (IEEE, 2015)
41. H. Lei, J. Choi, A. Janin, G. Friedland, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. User verification: matching the uploaders of videos across accounts (IEEE, 2011), pp. 2404–2407
42. S. Chu, S. Narayanan, C. C. J. Kuo, Environmental sound recognition with time-frequency audio features. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1142–1158 (2009)
43. B. Mathieu, S. Essid, T. Fillon, J. Prado, G. Richard, in *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Yaafe, an easy to use and efficient audio feature extraction software, (Utrecht, 2010)
44. F. Metze, S. Rawat, Y. Wang, in *Multimedia and Expo (ICME), 2014 IEEE International Conference on*. Improved audio features for large-scale multimedia event detection (IEEE, 2014), pp. 1–6
45. A. F. Smeaton, P. Over, W. Kraaij, in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. Evaluation campaigns and TRECVid (Association for Computing Machinery, New York, 2006), pp. 321–330
46. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al., Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
47. S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference On*. CNN architectures for large-scale audio classification (IEEE, 2017), pp. 131–135
48. A. Neviarouskaya, H. Prendinger, M. Ishizuka, in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. Sentiful: generating a reliable lexicon for sentiment analysis, (2009), pp. 1–6
49. A. Datta, M. Shah, N. D. V. Lobo, in *Pattern Recognition, 2002. Proceedings. 16th International Conference On, vol 1*. Person-on-person violence detection in video data (IEEE, 2002), pp. 433–438
50. C. Guastavino, The ideal urban soundscape: investigating the sound quality of french cities. Acta Acustica U. Acustica. **92**(6), 945–951 (2006)