"Are You Playing a Shooter Again?!" Deep Representation Learning for Audio-Based Video Game Genre Recognition

Shahin Amiriparian, *Student Member, IEEE*, Nicholas Cummins, *Member, IEEE*, Maurice Gerczuk, Sergey Pugachevskiy, Sandra Ottl, and Björn Schuller, *Fellow, IEEE*

***Abstract*—In this paper, we present a novel computer audition task: audio-based video game genre classification. The aim of this study is threefold: 1) to check the feasibility of the proposed task; 2) to introduce a new corpus: The Game Genre by Audio + Multimodal Extracts (G^2AME), collected entirely from social multimedia; and 3) to compare the efficacy of various acoustic feature spaces to classify the G^2AME corpus into six game genres using a linear support vector machine classifier. For the classification we extract three different feature representations from the game audio files: 1) Knowledge-based acoustic features; 2) DEEP SPECTRUM features; and 3) quantized DEEP SPECTRUM features using Bag-of-Audio-Words. The DEEP SPECTRUM features are a deep-learning-based representation derived from forwarding the visual representations of the audio instances, in particular spectrograms, mel-spectrograms, chromagrams, and their deltas through deep task-independent pretrained CNNs. Specifically, activations of fully connected layers from three common image classification CNNs, GoogLeNet, AlexNet, and VGG16 are used as feature vectors. Results for the six-genre classification problem indicate the suitability of our deep learning approach for this task. Our best method achieves an accuracy of up to 66.9% unweighted average recall using tenfold cross-validation.**

***Index Terms*—Audio classification, convolutional neural network (CNN), deep learning, game genre classification.**

This work was supported in part by the European Unions's Seventh Framework Programme under Grant 338164 (ERC StG iHEARu), and in part by the EU's Horizon 2020 Programme under Grant 688835 (DE-ENIGMA). *(Corresponding author: Shahin Amiriparian.)*

S. Amiriparianm is with the Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany and also with the Machine Intelligence and Signal Processing Group, Technische Universität München, 80333 Munich, Germany (e-mail: shahin.amiriparian@informatik.uni-augsburg.de).

N. Cummins, M. Gerczuk, S. Pugachevskiy, and S. Ottl are with the Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: nicholas.cummins@informatik.uni-augsburg.de; maurice.gerczuk@informatik.uni-augsburg.de; sergey.pugachevskiy@informatik.uni-augsburg.de; sandra.ottl@informatik.uni-augsburg.de).

B. Schuller is with the Chair of Embedded Intelligence for Health Care & Wellbeing, Univeristy of Augsburg, 86159 Augsburg, Germany and also with the GLAM – the Group on Language, Audio and Music, Imperial College London, SW7 2AZ London, U.K. (e-mail: schuller@tum.de).

I. Introduction

VIDEO games, despite being a relativity new form of media, are a highly important and influential aspect of popular culture. Moreover, the video game market has become an established and ever-growing global industry. In terms of total revenue generated, the video game industry rivals the more traditional entertainment industries such as music, film, and television [1], [2]. Due to this popularity, video games have become the subject of an active and expanding research field, which includes many diverse disciplines such as sociology, cultural studies, philosophical and behavioral psychology, and informatics [1].

Genre classification consists of grouping different objects into categories; for video games these categories are typically derived according to gameplay and interactivity [3], [4]. The study of video game genre differs considerably from film or literary genre study, because of the active participation of the gamer (audience), through the surrogate player-character who acts within the games diegetic world [5]. Research into the classification of video games into specific genres is of particular interest in both the video game industry as well as academia [3], [4].

Video games are an interactive audio-visual and tactile medium, with the soundscape playing a key role in the overall gaming experience [6]. The development of game audio can be considered as the result of a series of technological, economic, ideological, social, and cultural pressures [7]. Further, elements such as genre and audience expectations often constrain gameplay audio [7]. Aspects such as space, time, narrative, and gameplay dynamics all play an influencing role in this regard.

Key audio events in games include: *Vocalizations* of game characters, *sound effects* relating to gameplay, *ambient effects* relating to atmosphere, and the *music* of the game [1]. The mix of these events within a particular game depends on gameplay mechanics and is highly related to the genre [6]. For example, action and shooting games such as the *Call of Duty* series will contain loud, sudden events including punches and gunshots; sports games such as the *FIFA* football series tend to have ongoing commentary voice-overs; racing games such as the *Forza* series contains a substantial amount of car noises including heavy accelerations and screeching brakes; finally, classic arcade games, for example, *Super Mario* or *Sonic the*

Hedgehog have *chiptune* (8-bit synthesized electronic music) based accompanying sounds.

In the era of machine learning, the field of computational and artificial intelligence (CI/AI) in games is rapidly developing and gaining more attention from the scientific community [8]–[10]. Artificial neural networks (ANNs), support vector machines (SVMs), and deep neural network based classifiers have been shown to be effective for improving the logical and rational decision-making processes in different game genres, and analysing the training behavior in serious games [11], [12].

Given the links between the audio content of a video game and its genre [1], [6] and the growing influence of CI/AI in game development [11], [12], this paper explores different machine learning based acoustic detection paradigms for the task of Video Game Genre Classification. To the best of the authors' knowledge, this is the first time such a study has been undertaken. As well as being a challenging machine learning task, this work has many potential real-world applications.

1) Development of a remote and unobtrusive tool to automatically monitor game usage. Such a tool could allow parents to better track their child's video game habits monitoring total play-time and, checking if the game is age appropriate [13]. Any such tool should of course be developed under a clear ethical framework ensuring that it has the goal of monitoring strictly for health and wellness purposes whilst maintaining and protecting privacy [14].
2) General activity monitoring, e.g., in smart homes. An analogous example are apps that collect TV-viewing data for advertisers using a smartphone's microphone; according to a recent report in *The New York Times*,[1] there are at least 250 of such apps currently on the market.
3) First step toward a game *SHAZAM* which can, e.g., identify game genres, game's name, game walkthroughs, or game tutorials based on a short audio sample played and using the microphone on a device.
4) As a simple objective aid for game designers to boost the decision making process for choosing the proper game sound of new games and check if the game has a suitable soundscape for a particular genre.
5) Aiding the automatic generation of game genre-specific music tracks.
6) Monitoring of games on YouTube and other social media platforms. As well as aiding the automatic segmentation of gameplay clips posted on social media into semantically meaningful chunks.
7) Automatic social media-based game retrieval system.

Our contribution in this paper includes the introduction of our video game corpus,[2] Game Genre by Audio + Multimodal Extracts (G^2AME), a collection of 1566 audio clips taken from 300 different video games across and grouped into six genres. These data were collected using our Cost-efficient Audio-visual Acquisition via Social-media Small-world Targeting (CAS^2T) toolkit for efficient large-scale big data collection from the video sharing website YouTube [15].

The classification paradigms explored for the G^2AME corpus have been adopted from techniques successfully used in other acoustic-detection tasks. Our baseline system is based on acoustic feature sets which are extracted from audio clips of different video games using our *openSMILE* toolkit [16]. These feature sets are primarily used in computational paralinguistics based detection tasks [17]. However, they have also been used for movie genre classification [18], and music genre classification [19], [20].

We compare the efficacy of these standard acoustic feature sets with state-of-the-art feature representations, namely DEEP SPECTRUM features and their quantised representation. DEEP SPECTRUM features are derived by first forwarding visual representations of audio data through deep convolutional neural networks (CNNs), such as AlexNet [21], GoogLeNet [22], and VGG16 [23]. These CNNs have been pretrained on images from the ImageNet data set,[3] including categories, such as birds, flowers, fruits, furniture, tools, vehicles, and persons for image classification. The activations of the fully connected layers of these CNNs, i.e., DEEP SPECTRUM features can then be used as an audio feature representation. DEEP SPECTRUM features have shown their versatility in a range of audio classification tasks, including snore sound detection [24], [25], audio-based sentiment analysis [26], acoustic scene classification [27], speech-based emotion detection [28], and autism severity detection [29].

Pretrained CNNs have been chosen for the deep feature extraction from audio data for the following reasons: First, the convolutional layers of CNNs are able to make strong assumptions with regard to locality of the pixel dependencies [21], [30]–[32]. Furthermore, because of the richness of the time-frequency information in (mel-)spectrograms, local structures relating to properties such as loudness, pitch, rhythm, and spectral energy distribution are inherently present, and are in turn readable for the CNNs. This is verified by the strong performance of the DEEP SPECTRUM features in a range of audio tasks, e.g., [24], [26]–[29]. Second, if fine-tuning or training a new deep learning model are performed on our G^2AME data set (in which data quantity is limited), we would have a high risk of over-fitting to the training data.

In order to quantize the DEEP SPECTRUM features, we also compute a Bag-of-Audio-Words (BOAW) representation, herein denoted as Bag-of-Deep-Features (BODF). BODF representations are considered more robust than raw features; the quantization step can be seen to be quasifiltering against small amounts of noise in a data set [26], [33].

The rest of this paper is laid out as follows. Section II introduces the G^2AME corpus. Section III outlines our machine learning methods for extracting acoustic and DEEP SPECTRUM features from the audio files. The classification experiments and the evaluation metrics are outlined in Section IV. The obtained results are given in Section V, before concluding the paper in Section VI.

[1] Retrieved January 05, 2018, from https://nyti.ms/2E7Iins

[2] Text (csv) files containing the YouTube video IDs for the videos in each class of the G^2AME corpus can be downloaded from the following link: https://www.dropbox.com/s/twpq1cnvvhqbakn/youtubeIDs-game-dataset.zip?dl=0.Password:nJ-mP10U

[3] Summary and statistics of the ImageNet data set: http://image-net.org/about-stats

TABLE I
THE DISTRIBUTION OF AUDIO CLIPS IN THE G^2 AME CORPUS ACROSS THE SIX GENRES

Genre	**Videos** (5s Clips)
ACS	258 (3096)
ARP	205 (2460)
FHT	330 (3960)
RCG	296 (3552)
SPT	266 (3192)
SWB	211 (2532)
$\sum$	1 566 (18792)

The number of 5-s clips per genre is denoted in parentheses.

II. G^{2}AME CORPUS

The G^{2}AME data set was collected directly from YouTube. Using CAS^{2}T, our complex network analyser toolkit [15] we identified and downloaded 1566 unique gameplay videos representing 300 individual games. The data collection process was particularly time consuming, due in part to two main limitations: First, while using YouTube API it is only possible to iterate or process a given finite number of videos per day, and Second, a lot of YouTube game videos had a commentator voice included which was talking during the gameplay. Hence, we had to thoroughly check all targeted videos to ensure they contained gameplay audio only.

The downloaded videos were then converted from their original *.mp4* or *.webm* video formats into 16-kHz wav files cut into chunks of one minute in length. The total net playtime of G^{2}AME is 26 hours of gameplay. Further, each clip is cut to 12 individual five second chunks which are later used as a basis for feature extraction and classification in the non-(BoDF) systems. This results in a total of 18 792 audio chunks (cf. Table I).

Using the popular online shopping platform *Amazon*[4] as a guide, we categorized the games and according audio clips into six different genre groups.

1) *Action or Shooter* (ACS) games; 258 instances picked from games such as *Battlefield 1, Assassin's Creed, Dark Souls, Diablo*, or *Call of Duty*.
2) *Arcade or Platform* (ARP) games; 205 instances picked from games such as *Sonic the Hedgehog, Donkey Kong, Golden Axe, Pac-Man*, or *Super Mario Brothers*.
3) *Fighting* (FHT) games; 330 instances picked from games such as *Mortal Kombat, Street Fighter*, or *Tekken*.
4) *Racing* (RCG) games; 296 instances picked from games such as *Forza, Gran Turismo*, or *Need For Speed*.
5) *Sports* (SPT) games; 266 instances picked from games such as *FIFA, NBA, MLB, Pro Evolution*, or *WWE2*.
6) *Simulation or World Building* (SWB) games; 211 instances picked from games such as *Age of Empires, Minecraft, Tropico, Warcraft*, or *The Sims*.

Two different versions of the same game are treated as one game, e.g., the football games *FIFA 16* and *FIFA 15* are both considered examples of the *FIFA* game. Each genre contains clips from 50 distinct video games. For our cross-validation (CV) scheme, each of the tenfolds contains the instances of five distinct games from every genre.[5] This provides "game-independence" ensuring that our machine learning algorithms do not focus on recognizing specific games instead of their respective genres.

As already mentioned, the work presented in this manuscript focuses on the recognition of game genre using the audio modality only. We have focused on this strictly audio approach for the following reasons.

1) For practical use in a game monitoring application; the genre can be recognized from distance, such as in a personal assistant device, with no need to see or analyse the screen content. Furthermore, audio processing, in general, is often considered more lightweight than visual processing and hence is potentially better suited for real-time classification in embedded devices.
2) As discussed in the introduction, audio is an essential part of video game experience that helps to enhance the gaming experience. In this regard, music, which is not visible, plays a key role in establishing atmospheric difference between various game genres. Audio also provides instant feedback to the player's inputs, such as shooting a gun. This is an important factor to get a better analysis of player's gaming behavior.
3) Classification of the game genres is potentially more reliable from the audio modality. For example, role playing games, such as *Dark Souls, Witcher 3*, or *Fallout 4* are often visually diverse making it harder to infer the genre. Audio also gives cues about the visually invisible objects, monsters, animals, or persons in gameplay (e.g., a person behind a wall or a monster hidden in bushes). Obtaining such information can improve the performance of a game analysis toolkit.

III. FEATURE REPRESENTATIONS

Before starting to classify the game genres (cf. Section IV), we need to extract feature representations from each audio instance in the G^{2}AME corpus. In this section, we discuss three kinds of feature representations: 1) acoustic features; 2) DEEP SPECTRUM features; and 3) BoDF, i.e., the quantized representation of DEEP SPECTRUM features. We only quantized the best chunk-level features. In addition, quantized DEEP SPECTRUM features have been shown to be effective in related audio classification tasks [26], [34].

A. Acoustic Feature Sets

In order to better understand the nature of the G^{2}AME corpus and the advantages of the proposed classification approaches, we also compared the performance of our deep learning system based on DEEP SPECTRUM features with two conventional, expert-designed, acoustic feature sets used for the INTERSPEECH 2009 Emotion Challenge (IS09) [35] and the INTER-

[4]www.amazon.com

[5]For reproducibility, the make up of the fold will be included in the data set release.

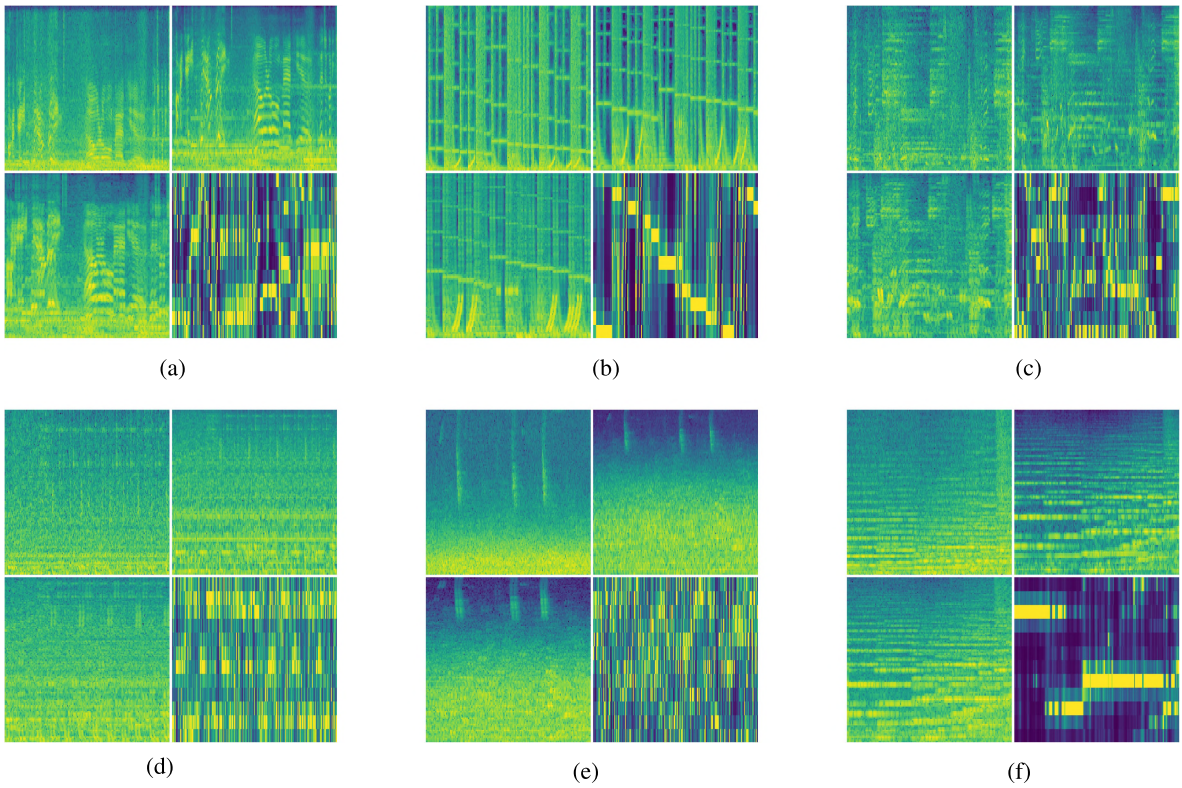


Fig. 1. Example spectrograms, mel-spectrograms, delta-mel-spectrograms and chromagrams (left-to-right and top-to-bottom) of audio samples contained in the six different classes of the G^{2}AME corpus. (a) *Alice: Madn. Ret.* (*Action or Shooter*). (b) *Dr. Robotnik* (*Arcade or Platform*). (c) *AquaPazza* (*Fighting*). (d) *Daytona USA* (*Racing*). (e) *Cricket* (*Sports*). (f) *Cossacks* (*Sim. or World Building*).

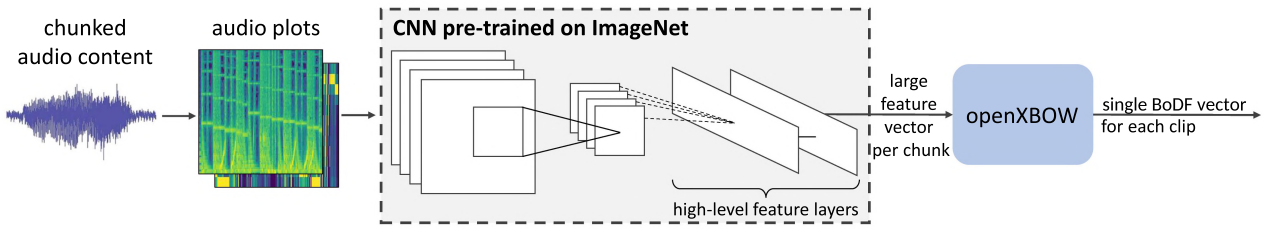


Fig. 2. Illustration of the proposed deep feature extraction method. A detailed account of the procedure is given in Sections III-B and III-C. Figure adapted from [34].

SPEECH 2010 Paralinguistic Challenge (IS10) [36], with 384 and 1582 features, respectively.

For full information on the extraction and formation of these feature sets, the interested reader is referred to the corresponding references as well as to [37].

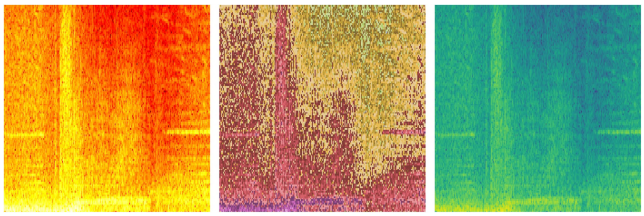
B. *Deep Spectrum Features*

We use a state-of-the-art system based on CNN image descriptors which extracts the Deep Spectrum features from plots of audio data, in particular spectrograms, mel-spectrograms, chromagrams, and their deltas or temporal transitions (cf. Section III-B1).

To highlight the audio similarities and differences that potentially exist between the game genres, the visual representations of audio samples contained in the six different classes taken from an exemplar game within each genre are shown in Fig. 1.

The basic system architecture is shown in the left part of Fig. 2. The features are extracted from audio data as follows. First, suitable representations are created using the audio and music analysis library *librosa* [38]. For our experiments on the G^{2}AME corpus, we extended the selection of audio representations to include mel-spectrograms and chromagrams apart from the standard spectrograms, since they have been successfully applied for various audio-based classification tasks [39]–[41].

These representations are then further transformed to images by creating color mapped plots. The second step consists of feeding these plots to CNNs pretrained on ImageNet [42] and extracting the activations of a specific fully-connected layer as large feature vectors. These features, denoted as Deep



(a) (b) (c)

Fig. 3. Sample spectrogram taken from the action/role-playing game Dark Souls displayed using three different color maps. (a) Hot. (b) Vega20b. (c) Viridis.

TABLE II
OVERVIEW OF THE ARCHITECTURAL SIMILARITIES AND DIFFERENCES BETWEEN THE TWO CNNS USED FOR THE EXTRACTION OF DEEP SPECTRUM FEATURES, ALEXNET AND VGG16

AlexNet	VGG16
input: RGB image	
1×conv size: 11; ch: 96; stride: 4	2×conv size: 3; ch: 64; stride: 1
maxpooling	
1×conv size: 5; ch: 256	2×conv size: 3; ch: 128
maxpooling	
1×conv size: 3; ch: 384	3×conv size: 3; ch: 256
	maxpooling
1×conv size: 3; ch: 384	3×conv size: 3; ch: 512
	maxpooling
1×conv size: 3; ch: 256	3×conv size: 3; ch: 512
maxpooling	
fully connected *fc6* 4 096 neurons	
fully connected *fc7* 4 096 neurons	
fully connected 1 000 neurons	
output: soft-max of probabilities for 1 000 object classes	

conv denotes convolutional layers and *ch* stands for channels. The table is adapted from [24].

SPECTRUM features are a high-level representation of the plots created from low-level audio features. All deep spectrum features were extracted using our purpose built toolkit.[6] Details about the used low-level audio features and pretrained CNN networks are described in the following subsections.

1) Audio Plots: In addition to standard spectrograms which are computed from Hanning windows of width 256 and overlap of 128 samples, we also use mel-spectrograms and chromagrams. The Hanning window helps to preserve both the frequency resolution and the amplitude of a signal. Mel-Spectrograms are computed from the log-magnitude spectrum by dimensionality reduction using a mel-filter. We use 128 filter banks equally spaced on the mel-scale defined as follows:

$$f_{\text{mel}} = 2595 \cdot \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \quad (1)$$

where f_{mel} is the resulting frequency on the mel-scale computed in mels and f_{Hz} is the normal frequency measured in hertz. The mel-scale is based on the frequency response of the human ear which has better resolution at lower frequencies. We also display the mel-spectrogram on this scale. Chromagrams are a mapping of the spectrogram bins into structures based around pitch relationships as defined by the western music tonality system. Compared to spectrograms, chroma features relate to the 12-pitch classes defined in the 12 tone equal temperament which are represented by the pitch spelling attributes: *C, C#, D, D#, E, F, F#, G, G# A, A#, and B*. As video games are often accompanied by scores or soundtracks, we hope that chromagrams capture some genre specific musical characteristics. Different ways of computing chroma features exist [43], we use the default implementation provided by the *librosa* Python library. Further, we compute the first-order derivatives (deltas) of the mel-spectrograms and chromagrams to incorporate more of the dynamics of the underlying features.

For the spectrogram plots, we use three different color mappings: *viridis*, *hot*, and *Vega20b*. These color mappings are shown in Fig. 3. It is during testing (cf. Section V) that we identify the optimal color map for the different spectral and chroma feature spaces.

2) CNN-Descriptors: To form suitable feature representations from the plots described in Section III-B1, three different ImageNet-trained CNNs are used as feature extractors.

AlexNet's and VGG16's architectures are compared in Table II. Here, we use the activations of the 4096 neurons on the second fully connected layer—denoted as *fc7*—as features because of their robustness for audio classification tasks [24], [26], [28]. The third network architecture—GoogLeNet—differs substantially from these two due to its so-called inception modules which concatenate convolutional filters of different size (cf. Fig. 4). These are then stacked with intermittent pooling layers and finally, a fully connected layer with 1000 neurons is used for ImageNet classification. Here, we apply the last pooling layer as feature vector. This results in feature vectors of size 1024.

For each of these CNN descriptors, we first evaluate the suitability of each color map on standard spectrograms and then use the best performing color map for the remaining experiments.

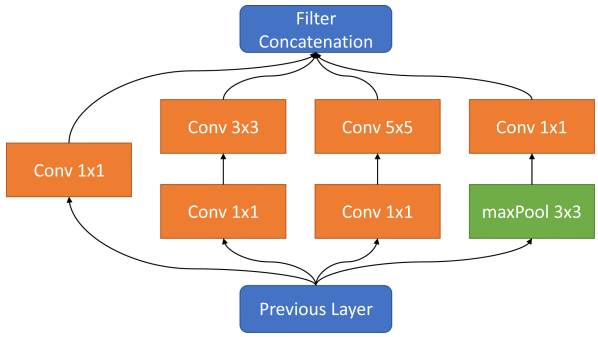


Fig. 4. Inception module used in the GoogLeNet architecture. Small 1 × 1 convolutions reduce dimensionality and filters of different patch sizes are concatenated to combine information found at different scales.

[6] https://github.com/DeepSpectrum/DeepSpectrum

C. Bag-of-Deep-Features

In order to form BoDF we combine the BoAW representation with the DEEP SPECTRUM features using OPENXBOW, our open-source toolkit for the generation of bag-of-words representations [44]. BoAW involves generating a fixed length histogram representation of an audio clip. This is achieved by first identifying a set of "deep audio words" from some given training data, and then quantizing (bagging) the original feature space, with respect to the generated codebook, to form the histogram representation. The histogram shows the frequency of each identified deep audio word in a given audio instance [33], [44]–[46]. It is worth noting, that the audio words do not represent words in their semantic meaning, but rather fragments of the audio signal defined by features [45]. The codebook can be the result of, e.g., a clustering algorithm [47] or a random sampling of low-level descriptors [48]. The histogram finally describes the distribution of the codebook vectors over the whole audio segment [44].

As shown in Fig. 2, we first extract a DEEP SPECTRUM representation for each five second chunk, then we bag them (12 per audio file, cf. Section II), to form a clip-level representation.

We normalize the features to $[0, 1]$ and random sample a codebook with fixed size from the training partition. Afterwards, each input feature vector is applied a fixed number of its closest vectors from the codebook. We then use logarithmic term-frequency weighting to the resulting histograms. The size of the codebook and the number of assigned codebook words (*cw*) are optimized with $size \in \{100, 200, 500, 1\,000\}$, $cw \in \{1, 10, 25, 50, 100, 200\}$.

IV. CLASSIFIERS AND EVALUATION METRICS

In order to predict the class labels for the audio instances in the G^{2}AME corpus, we train a linear SVM classifier on the extracted feature representations. The evaluation measure is unweighted average recall (UAR). We use UAR, as this measure gives equal weight to all classes of our G^{2}AME corpus and is accordingly more suitable than a weighted metric (e.g., accuracy) for our data set which has slightly imbalanced class distribution (cf. Table I).

We use the open-source linear SVM implementation provided in the scikit-learn machine learning library [49]. Feature standardization, i.e., subtracting the mean and dividing by the standard deviation is applied to the openSMILE feature sets. For the DEEP SPECTRUM features, both standardization and normalization have been found to negatively impact classifier's performance. We use the built-in balancing option of the SVM classifier to counteract the slight imbalance of our data set. The classifiers complexity parameter is optimized in ten steps, equally spaced on a logarithmic scale between 10^{-9} and 10^{0}.

V. RESULTS

An extensive series of experiments has been conducted to evaluate the performance of the extracted feature representations (cf. Section III) using the proposed classifiers (cf. Section IV).

First, we evaluate the performance of the acoustic feature sets extracted with openSMILE (cf. Section III-A). We then obtain the classification results for the DEEP SPECTRUM features (cf. Section V-B). Afterwards, we evaluate the effect of quantizing the best performing feature plots for each CNN descriptor (cf. Section V-C). In Section V-D, we perform model (late) fusion on the DEEP SPECTRUM and acoustic feature representations. Finally, we compare our best classification result with human performance (cf. Section V-E).

TABLE III
PERFORMANCE OF THE SVM CLASSIFIER USING DIFFERENT OPENSMILE AUDIO FUNCTIONALS ON THE G^{2}AME CORPUS

Acoustic feature set	C	Tenfold CV
IS09	10^{-6}	49.6 (49.2 ± 4.1 %)
IS10	10^{-6}	55.2 (54.7 ± 4.5 %)

UAR on concatenated folds is provided. Mean and standard deviation are reported in parentheses.

We perform statistical T-tests, comparing the results from acoustic feature sets (cf. Table III) with the best DEEP SPECTRUM results (cf. Table V) and the overall best performance (cf. Table VI) in a pairwise fashion to determine if they are statistically different. We reject the null-hypothesis at a significance level of $p < 0.05$. For each comparison, the p-values can be found in Table VIII. We checked the results of the cross-validation for the normality using a Shapiro–Wilk test [50], [51].

A. Acoustic Feature Sets

The two sets of audio functionals extracted with openSMILE were also evaluated using tenfold CV with linear SVM classifier. We observe that the larger feature set (IS10) with 1582 features outperforms the smaller one (IS09) with 384 features (cf. Table III) reaching a maximum UAR of 55.2%. We also show that there is statistically significant difference between these two features (cf. Table VIII).

B. DEEP SPECTRUM Features

The comparison of different feature plots as basis for DEEP SPECTRUM extraction (cf. Table V) indicates that chromagrams—in contrast to their relevance to music analysis—do not provide suitable input for ImageNet pre-trained CNNs. We explain this partially by the inherent "unnatural" look of these plots, leading to an inability to extract informative structural features using an Image CNN. Applying mel-spectrograms on the other hand, slightly improves performance for both AlexNet and GoogLeNet. Using the deltas decrease the performance for all CNNs. The overall best performance is achieved by AlexNet features extracted from mel-spectrograms with a UAR of 59.9%.

C. Bag-of-Deep-Features

For the BoDF representations, we chose the best performing feature plots for each CNN, i.e., mel-spectrograms for both AlexNet and GoogLeNet and regular spectrograms for VGG16. We optimized the BoDF parameters codebook size and number of assigned codebook words *cw* (cf. Section III-C) as well as the SVM's complexity parameter. The results in Table VI show a maximum UAR of 66.9% achieved by BoDF with a codebook size of 500 and *cw* of 25 formed from DEEP SPECTRUM

TABLE IV
PERFORMANCE OF THE SVM CLASSIFIER USING DEEP SPECTRUM FEATURES EXTRACTED FROM SPECTROGRAM PLOTS WITH DIFFERENT COLOR MAPS BY THREE CNN-DESCRIPTORS ON THE G^2 AME CORPUS

	Colour map / CNN-descriptor	AlexNet	GoogLeNet	VGG16
SVM	hot	58.1 (57.5 ± 3.2 %)	55.3 (54.7 ± 4.6 %)	58.8 (57.9 ± 4.0 %)
	Vega20b	55.2 (54.7 ± 3.8 %)	52.3 (51.5 ± 4.1 %)	55.7 (54.8 ± 4.2 %)
	viridis	58.6 (57.9 ± 3.7 %)	54.7 (54.0 ± 3.8 %)	58.3 (57.6 ± 3.1 %)

UAR on concatenated folds is provided. Mean and standard deviation are reported in parentheses.

TABLE V
PERFORMANCE OF THE SVM CLASSIFIER USING BEST PERFORMING DEEP SPECTRUM FEATURES ON THE G^2 AME CORPUS FROM TABLE IV

% UAR	AlexNet (viridis)	GoogLeNet (hot)	VGG16 (hot)
spectrograms	58.6 (57.9 ± 3.7 %)	55.3 (54.7 ± 4.6 %)	**58.8** (57.9 ± 4.0 %)
mel-spectrograms	**59.9** (59.2 ± 4.7 %)	**58.3** (57.6 ± 4.1 %)	58.7 (57.8 ± 4.2 %)
Δmel-spectrograms	56.5 (55.7 ± 4.9 %)	54.8 (53.9 ± 4.5 %)	55.2 (54.3 ± 4.6 %)
chroma	46.2 (45.4 ± 4.1 %)	46.2 (45.3 ± 3.1 %)	47.6 (46.6 ± 2.9 %)
Δchroma	41.1 (40.3 ± 3.6 %)	38.7 (37.9 ± 2.4 %)	41.1 (40.3 ± 2.9 %)

UAR on concatenated folds is provided. Mean and standard deviation are reported in parentheses.

TABLE VI
PERFORMANCE OF THE SVM CLASSIFIER USING BODF REPRESENTATIONS OF THE BEST DEEP SPECTRUM FEATURES ON THE G^2 AME CORPUS

CNN-descriptor	S	cw	C	Tenfold CV
AlexNet mel-spectrograms (viridis)	1 000	200	10^{-6}	66.7 (66.0 ± 4.3 %)
	1 000	100	10^{-5}	66.6 (64.7 ± 4.7 %)
GoogLeNet mel-spectrograms (hot)	500	25	10^{-5}	**66.9** (66.3 ± 4.0 %)
	500	100	10^{-6}	66.3 (64.8 ± 4.2 %)
VGG16 spectrograms (hot)	1 000	100	10^{-6}	65.1 (64.6 ± 6.4 %)
	1 000	50	10^{-6}	64.8 (64.2 ± 5.5 %)

S: codebook size; cw: number of assigned codebook words; C: SVM classifier's complexity. UAR on concatenated folds is provided. Mean and standard deviation are reported in parentheses.

features and GoogleNet mel-spectrograms as CNN descriptor. We also show that there is statistically significant difference between GoogLeNet BoDF system and the other tested approaches (cf. Table VIII). A confusion matrix for this system is displayed in Fig. 5(b).

D. Fusion Experiments

In addition to the quantization method employed in Section V-C, we also perform late fusion on the different chunk-level feature representations. We fuse the best performing DEEP SPECTRUM features with each of the three audio functional feature sets and also all four of them together. Note that, we also attempted an early fusion, however initial analysis revealed this approach was not suitable.

Our late fusion scheme uses the trained and optimized SVM models obtained in Section V-B and Section V-A and combines their predictions data by majority vote. These results (cf. Table VII) further indicate that the different feature sets are not complementary.

TABLE VII
PERFORMANCE OF LATE FUSION USING LINEAR SVM CLASSIFIER ON THE G^2 AME CORPUS

Late fusion	Tenfold CV
DEEP SPECTRUM + IS09	56.0 (55.4 ± 4.0 %)
DEEP SPECTRUM + IS10	58.3 (57.7 ± 4.2 %)
all features combined	57.8 (57.4 ± 4.5 %)

We employ a majority vote using the best individual models obtained during previous experiments. UAR on concatenated folds is provided. Mean and standard deviation are reported in parentheses.

E. Comparison With Human Performance

To gain perspective into how well our best classification approach performed (cf. Table VI) we conducted human classification tests through our browser-based crowdsourcing platform iHEARu-PLAY [52]. iHEARu-PLAY is a modular, browser-based crowdsourcing platform and is publicly available for users.[7] For the perception task, we presented the human raters with one 5-s clip (picked at random) for each file. A total of 12 individuals (seven male, five female, average age 27.3, nonprofessional gamers) completed the full classification task on iHEARu-PLAY platform. The average per rater decision time was 3.31 s.

[7]https://www.ihearu-play.eu

TABLE VIII
p-VALUES FOR T-TEST SCORES COMPARING CROSS-VALIDATION RESULTS OF DIFFERENT CONFIGURATIONS

	GoogLeNet BoDF (66.9)	AlexNet (59.9)	VGG16 (58.8)	IS09 (49.6)	IS10 (55.2)
GoogLeNet BoDF (66.9)	1.0	$4 \cdot 10^{-4}$	$4 \cdot 10^{-4}$	$5 \cdot 10^{-8}$	$2 \cdot 10^{-5}$
AlexNet (59.9)	-	1.0	0.522	$1 \cdot 10^{-4}$	$4.9 \cdot 10^{-2}$
VGG16 (58.8)	-	-	1.0	$2 \cdot 10^{-4}$	0.128
IS09 (49.6)	-	-	-	1.0	0.014
IS10 (55.2)	-	-	-	-	1.0

Except for AlexNet vs. VGG16 and VGG16 vs. IS10 the difference between other feature sets is statistically significant.

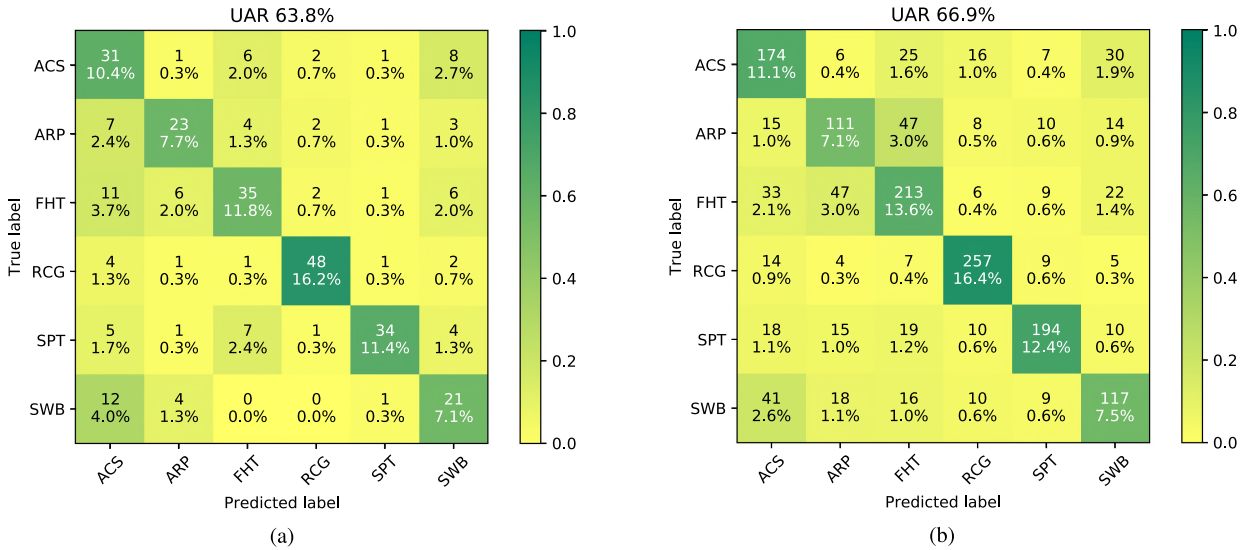


Fig. 5. Confusion matrices from the human performance and our best classification result. (a) Confusion Matrix from classification labels for 297 audio files gained by the best human rater. (b) Confusion Matrix from classification labels for the 1 566 audio files in the G^{2}AME data set gained by our best BoDF system with $cw = 25$ and $S = 500$.

Taking the average of the UAR scores gives us an overall human rater UAR of 59.7%. The best human performance was 63.8% UAR which is 3.1 percentage points less than the best classification result. The confusion matrix for this prediction is given in Fig. 5(a).

While the raters' scores give us a reference baseline on which to gauge our machine learning approaches, it should be noted that due to the differing amounts of training, as well as the cross-validation paradigm used in the machine learning approaches, such a comparison should not be considered like-for-like. Furthermore, our machine learning approaches were trained on a large amount of data (cf. Table I) which the raters did not have access to. However, given the prevalence of video games in today's society [1], [2], one cannot say that the raters received no training. Further, the terminology associated with each genre label would have helped the raters form preconceptions of what a "typical" audio clip from a particular genre should sound like. Despite these factors, the stronger performance of our system indicates the suitability of using machine learning for the task of video game genre classification.

VI. CONCLUSION AND FUTURE WORK

Video games as a rapidly growing entertainment medium have a complex and evolving taxonomy. The associated genrefication receives considerable attention both in research and industry. In this regard, the work presented in this paper explored, for the first time, audio-based video genre classification. In Section I, we listed potential real-world applications of such a system. In Section II, we presented the novel G^{2}AME data set, which comprises 1566 unique gameplay samples taken from 300 individual video games. The samples were downloaded from YouTube using a graph based search toolkit [15], that exploits the small-world nature of YouTube's recommendation system to efficiently source data. In Section III, we then explored the efficacy of a range of acoustic event detection paradigms for the task, including state-of-the-art DEEP SPECTRUM features. Our results in Section V indicated that BoDF, a combination of DEEP SPECTRUM features and BoAW, are well suited to the task of audio-based game-genre classification. This system achieved the strongest UAR of 66.9% an improvement of 3.1 percentage points over humans performing the same task. We also performed statistical T-tests to compare the results obtained from various feature sets in a pairwise manner and showed that there is statistically significant differences between the overall best result and the results obtained from other proposed feature vectors (cf. Table VIII).

An in-depth analysis of the results reveals that *Racing* games, which normally feature a substantial amount of automotive noises, were the easiest to recognize. *Simulation and World*

Building games were the most difficult to analyse, having confusion with the *Action, Adventure, or Shooter* genre.

We have a wide-ranging set of future work plans in association with this task. First, we aim to greatly grow the G^2AME data set. We will verify and expand on the audio based classification task offered within this paper. We will also consider visual-based, linguistic-based (YouTube comments), and multimodal classification, which has shown to improve classifier accuracy in related tasks such as audio-visual based sports genre classification [53]. Given the complexity of video game labels, other future work will include using advanced clustering techniques, treating each video game as being associated with a set of genre labels, rather than our current method of assigning each a single label, in a multimodal framework. We speculate such an approach could aid the identification of new sets of video game genre descriptors and labels.

Acknowledgment

The authors would like to thank the reviewers for their thoughtful comments and efforts towards improving the manuscript.

References

[1] S. Egenfeldt-Nielsen, J. H. Smith, and S. P. Tosca, *Understanding Video Games: The Essential Introduction*, 3rd ed. New York, NY, USA: Routledge, 2016.

[2] A. Marchand and T. Hennig-Thurau, "Value creation in the video game industry: Industry economics, consumer benefits, and research opportunities," *J. Interact. Marketing*, vol. 27, no. 3, pp. 141–157, 2013.

[3] T. Apperley, "Genre and game studies: Toward a critical approach to video game genres," *Simulation Gaming*, vol. 37, no. 1, pp. 6–23, Mar. 2006.

[4] R. I. Clarke, J. H. Lee, and N. Clark, "Why video game genres fail: A classificatory analysis," *Games Culture*, vol. 12, no. 5, pp. 445–465, Jul. 2015.

[5] M. J. Wolf, "Genre and the video game," in, *The Medium of the Video Game*. Austin, TX, USA: Univ. Texas Press, 2001.

[6] K. Collins, *Playing with Sound : A Theory of Interacting with Sound and Music in Video Games*, 1st ed. Cambridge, MA, USA: MIT Press, 2013.

[7] K. Collins, "Game sound," *An Introduction History, Theory, Practice Video Game Music Sound Design.* Cambridge, MA, USA: MIT Press, 2008.

[8] S. Risi and J. Togelius, "Neuroevolution in games: State of the art and open challenges," *IEEE Trans. Games*, vol. 9, no. 1, pp. 25–41, Mar. 2017.

[9] G. N. Yannakakis and J. Togelius, "A panorama of artificial and computational intelligence in games," *IEEE Trans. Games*, vol. 7, no. 4, pp. 317–335, Dec. 2015.

[10] S. M. Lucas, "Computational intelligence and AI in games: A new ieee transactions," *IEEE Trans. Games*, vol. 1, no. 1, pp. 1–3, Mar. 2009.

[11] M. Frutos-Pascual and B. G. Zapirain, "Review of the use of AI techniques in serious games: Decision making and machine learning," *IEEE Trans. Games*, vol. 9, no. 2, pp. 133–152, Jun. 2017.

[12] M. C. Gombolay, R. Jensen, and S. H. Son, "Machine learning techniques for analyzing training behavior in serious gaming," *IEEE Trans. Games*, to be published.

[13] C. Steinkuehler, "Parenting and video games," *J. Adolescent Adult Literacy*, vol. 59, no. 4, pp. 357–361, 2016.

[14] IEEE Global Initiative, "Ethically aligned design," Accessed: Oct. 1, 2018. [Online]. Available: https://ethicsinaction.ieee.org

[15] S. Amiriparian *et al.*, "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," in *Proc. 7th Biannu. Conf. Affective Comput. Intell. Interact.*, San Antonio, TX, USA, Oct. 2017, pp. 340–345.

[16] F. Eyben *et al.*, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, Barcelona, Spain, Oct. 2013, pp. 835–838.

[17] B. Schuller *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, San Francisco, CA, USA, Sep. 2016, pp. 2001–2005.

[18] A. Austin *et al.*, "Characterization of movie genre based on music score," in *Proc. 35th IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 421–424.

[19] A. Rosner *et al.*, "Influence of low-level features extracted from rhythmic and harmonic sections on music genre classification," in *Man-Machine Interaction 3, Advances in Intelligent Systems and Computing*, A. Gruca, T. Czachrski, and S. Kozielski, Eds., Cham, Switzerland, Springer, 2014, vol. 242, pp. 467–473.

[20] A. Rosner, B. Schuller, and B. Kostek, "Classification of music genres based on music separation into harmonic and drum components," *Arch. of Acoust.*, vol. 39, no. 4, pp. 629–638, 2015.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 26th Ann. Conf. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[22] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1–9.

[23] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 568–576.

[24] S. Amiriparian *et al.*, "Snore sound classification using image-based deep spectrum features," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, Stockholm, Sweden, Aug. 2017, pp. 3512–3516.

[25] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. W. Schuller, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *J. Mach. Learn. Res.*, 2017, arXiv:1712.04382.

[26] S. Amiriparian, N. Cummins, S. Ottl, M. Gerczuk, and B. Schuller, "Sentiment analysis using image-based deep spectrum features," in *Proc. 2nd Int. Workshop Autom. Sentiment Anal. Wild, WASA ,held Conjunction 7th Biannu. Conf. Affective Comput. Intell. Interact.*, San Antonio, TX, USA, Oct. 2017, pp. 26–29.

[27] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. Detect. Classification Acoust. Scenes Events 2017 Workshop*, Munich, Germany, Nov. 2017, pp. 17–21.

[28] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proc. 25th ACM Int. Conf. Multimedia*, Mountain View, CA, USA, Oct. 2017, pp. 478–484.

[29] A. Baird *et al.*, "Automatic classification of autistic child vocalisations: A novel database and results," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, Stockholm, Sweden, Aug. 2017, pp. 849–853.

[30] K. Jarrett *et al.*, "What is the best multi-stage architecture for object recognition?" in *Proc. IEEE 12th Int. Conf. Comput. Vision*, 2009, pp. 2146–2153.

[31] A. Krizhevsky and G. Hinton, "Convolutional deep belief networks on cifar-10," to be published.

[32] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2004, vol. 2, Paper II-104.

[33] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, San Francsico, CA, USA, Sep. 2016, pp. 495–499.

[34] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, S. Pugachevskiy, and B. Schuller, "Bag-of-deep-features: Noise-robust deep feature representations for audio analysis," in *Proc. 31st Int. Joint Conf. Neural Netw.*, Rio de Janeiro, Brazil, Jul. 2018, pp. 1–7.

[35] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, Brighton, UK, Sep. 2009, pp. 312–315.

[36] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc*. Makuhari, Japan, Sep. 2010, pp. 2794–2797.

[37] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, (Springer Theses). Cham, Switzerland: Springer International Publishing, 2015.

[38] B. McFee *et al.*, *librosa 0.5.0*. (2017) Feb. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.293021

[39] S. Panwar, A. Das, M. Roopaei, and P. Rad, "A deep learning approach for mapping music genres," in *Proc. 12th Syst. Syst. Eng. Conf.*, 2017, pp. 1–5.
[40] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 171–175.
[41] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "Dcase 2016 acoustic scene classification using convolutional neural networks," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2016, pp. 95–99.
[42] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit*. Miami, FL, USA, Jun. 2009, pp. 248–255.
[43] M. Müller and S. Ewert, "Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features," in *Proc. 12th Int. Conf. Music Inf. Retrieval*, 2011, pp. 215–220.
[44] M. Schmitt and B. Schuller, "openxbow – introducing the passau open-source crossmodal bag-of-words toolkit," *J. Mach. Learn. Res.*, 2017, arXiv:1605.06778.
[45] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc*. Portland, OR, USA, 2012, pp. 2105–2108.
[46] S. Pancoast and M. Akbacak, "Softening quantization in bag-of-audio-words," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1370–1374.
[47] F. B. Pokorny, F. Graf, F. Pernkopf, and B. W. Schuller, "Detection of negative emotions in speech signals using bags-of-audio-words," in *Proc. Int. Conf. Affective Comput. Intell. Interact.*, 2015, pp. 879–884.
[48] S. Rawat, P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze, "Robust audio-codebooks for large-scale event detection in consumer videos." in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 2929–2933.
[49] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[50] J. Peat and B. Barton, *Medical Statistics: A Guide to Data Analysis and Critical Appraisal*. Hoboken, NJ, USA: Wiley, 2008.
[51] D. Öztuna, A. H. Elhan, and E. Tüccar, "Investigation of four different normality tests in terms of type 1 error rate and power under different distributions," *Turkish J. Med. Sci.*, vol. 36, no. 3, pp. 171–176, 2006.
[52] S. Hantke *et al.*, "iHEARu-PLAY: introducing a game for crowdsourced data collection for affective computing," in *Proc. 1st Int. Workshop Autom. Sentiment Anal. Wild (WASA 2015) held Conjunction with 6th Biannu. Conf. Affective Comput. Intell. Interact.*, Xi'an, China, Sep. 2015, pp. 891–897.
[53] J. Wang, C. Xu, and E. Chng, "Automatic sports video genre classification using pseudo-2d-hmm," in *Proc. 18th Int. Conf. Pattern Recognit*, Hong Kong, China, Aug. 2006, vol. 4, pp. 778–781.

Shahin Amiriparian (GSM'18) received the M.Sc. degree in electrical engineering and information technology from Technische Universität München (TUM), Munich, Germany. He started working toward the Ph.D. degree as a Researcher in the Machine Intelligence and Signal Processing Group at TUM, focusing his research on novel deep learning methods for audio processing.

From 2014 to 2017, he was a Doctoral Researcher with the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany and is currently pursuing his doctoral degree with the ZD.B Chair of Embedded Intelligence for Health Care and Well Being at the University of Augsburg, Augsburg, Germany. His research interests include machine learning methods, deep learning and representation learning for audio processing.

Nicholas Cummins (M'13) did his undergraduate degree at UNSW, Sydney, Australia, as a mature student, graduating with first class honors in 2011. He received the Ph.D. degree in electrical engineering from UNSW in February 2016. His Ph.D. investigated whether the voice can be used as an objective marker in the diagnosis and monitoring of clinical depression.

He is a habilitation candidate at the ZB.D Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany where he is involved in Horizon 2020 projects DE-ENIGMA, RADAR-CNS, and TAPAS. He has authored and coauthored regularly in the field of depression detection since 2011 and these papers have attracted considerable attention and citations. His current research includes areas of behavioral signal processing with a focus on the automatic multisensory analysis and understanding of different health states.

Maurice Gerczuk received the B.Sc. degree in computer science from the University of Passau, Passau, Germany. He is currently working toward the M.S. degree in computer science at the University of Augsburg, Augsburg, Germany.

He was a Student Research Assistant with the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany and is currently a Research Assistant at the ZD.B Chair of Embedded Intelligence for Health-Care and Wellbeing, University of Augsburg, Augsburg, Germany.

Sergey Pugachevskiy received the B.Sc. degree in computer science at the University of Passau, Passau, Germany. He is currently working toward the M.S. degree in computer science at the University of Augsburg, Augsburg, Germany.

He was a Student Research Assistant with the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany and is currently a Research Assistant at the ZD.B Chair of Embedded Intelligence for Health-Care and Wellbeing, University of Augsburg, Augsburg, Germany. His research interests include machine learning and big data.

Sandra Ottl received the B.Sc. degree in computer science from the University of Passau, Passau, Germany. She is currently working toward the M.S. degree in computer science at the University of Augsburg, Augsburg, Germany.

She was a Student Research Assistant with the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany and is currently a Research Assistant at the ZD.B Chair of Embedded Intelligence for Health-Care and Wellbeing. University of Augsburg, Augsburg, Germany, and works on the DE-ENIGMA project.

Björn Schuller (M'06, SM'15, F'18) received the Diploma in 1999, the Ph.D. degree in 2006, and the Habilitation and Adjunct Teaching Professorship in the subject area of signal processing and machine intelligence in 2012, all in electrical engineering and information technology from Technische Universität München, Munich, Germany.

He is a Professor of Artificial Intelligence in the Department of Computing, Imperial College London, London, U.K., and a Full Professor and head of the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg, Augsburg, Germany. He has authored and coauthored five books and more than 600 publications in peer reviewed books, journals, and conference proceedings leading to more than 20 000 citations (h-index = 68).