J. Abeßer, A. Ullah, S. Ziegler and S. Grollmisch,
"Human and Machine Performance in Counting Sound Classes
in Single-Channel Soundscapes"
J. Audio Eng. Soc., vol. 71, no. 12, pp. 860–872, (2023 December).
DOI: https://doi.org/10.17743/jaes.2022.0106

Human and Machine Performance in Counting Sound Classes in Single-Channel Soundscapes

JAKOB ABEßER,[1] ASAD ULLAH,[1]
(jakob.abesser@idmt.fraunhofer.de) (asad.ullah@idmt.fraunhofer.de)
SEBASTIAN ZIEGLER,[1] AND SASCHA GROLLMISCH[1, 2]
(sebastian.ziegler@idmt.fraunhofer.de) (sascha.grollmisch@idmt.fraunhofer.de)

[1] *Semantic Music Technologies, Fraunhofer Institute for Digital Media Technology (IDMT), Ilmenau, Germany*
[2] *Industrial Media Applications, Fraunhofer Institute for Digital Media Technology (IDMT), Ilmenau, Germany*

Individual sounds are difficult to detect in complex soundscapes because of a strong overlap. This article explores the task of estimating sound polyphony, which is defined here as the number of audible sound classes. Sound polyphony measures the complexity of a soundscape and can be used to inform sound classification algorithms. First, a listening test is performed to assess the difficulty of the task. The results show that humans are only able to reliably count up to three simultaneous sound sources and that they underestimate the degree of polyphony for more complex soundscapes. Human performance depends mainly on the spectral characteristics of the sounds and, in particular, on the number of overlapping noise-like and transient sounds. In a second step, four deep neural network architectures, including an object detection approach for natural images, are compared to contrast human performance with machine learning–based approaches. The results show that machine listening systems can outperform human listeners for the task at hand. Based on these results, an implicit modeling of the sound polyphony based on the number of previously detected sound classes seems less promising than the explicit modeling strategy.

0 INTRODUCTION

The human auditory system shows its impressive capabilities when listening to complex soundscapes, which range from scenarios with multiple active speakers (often referred to as cocktail party scenarios), music ensembles with several instruments, and everyday soundscapes with multiple static and moving sound sources. In many tasks such as audio source separation, speech recognition, and music transcription, algorithms benefit from prior information about the number of sound sources (sound polyphony).

In the auditory scene analysis model proposed by Bregman [1], the human auditory system groups audible sounds along frequency (simultaneous grouping) and time (sequential grouping). As a result of this grouping, sounds are either integrated into a single auditory perception, which is mentally assigned to a single sound source, or segregated into different auditory streams, which are assigned to individual sound sources. In dense soundscapes particularly, grouping errors can cause the auditory stream segregation to fail and overlapping sounds to be perceived as blended sounds, which complicates their classification.

The concept of polyphony can be interpreted from different perspectives, and its estimation is addressed in various research tasks. In the music domain, polyphony denotes the number of simultaneously sounding pitched sound events (tones) or temporal sequences thereof (melodic lines) [2]. A closely related task is music ensemble size estimation, which aims to estimate of the number of simultaneously active instruments [3]. In speech processing, a common task is to estimate the number of active speakers (speaker counting). In the computer vision domain, related research tasks are object counting [4] and face counting [5] in natural images.

In this study, the authors investigate the task of sound polyphony estimation (SPE) for everyday sounds. Salamon et al. define sound polyphony as the maximum number of overlapping sounds at any time in an audio recording [6]. As a disadvantage, this definition requires precise sound event annotations, which are ill-defined for sound classes with ambiguous start and end times. Therefore, the authors propose an alternative definition and measure sound polyphony as the number of unique sound classes audible in a short audio segment. This definition focuses on sound event tagging (SET), i.e., classifying audible sounds without precisely lo-

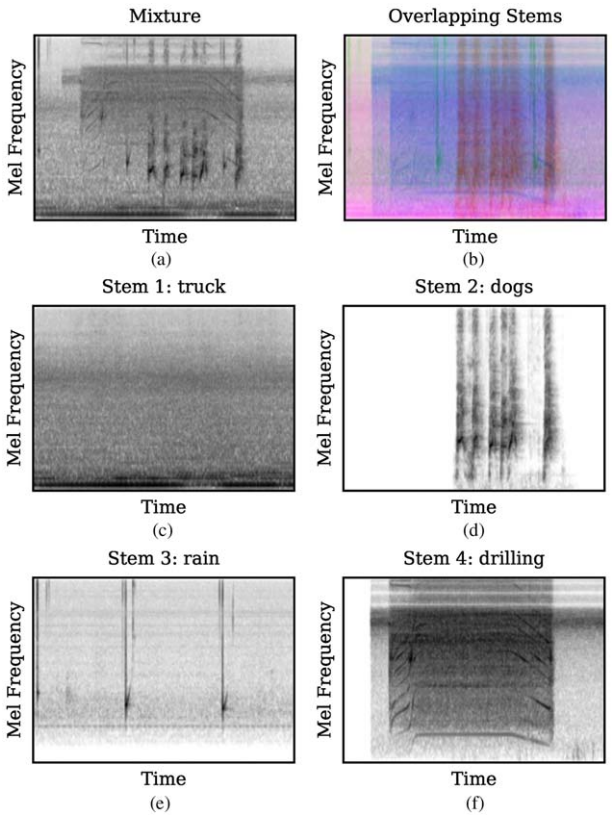

Fig. 1. Example of a polyphonic soundscape (USM evaluation set, `445.wav`, sound polyphony degree of $y_p = 4$). Log-magnitude–scaled Mel spectrograms are shown for the four underlying audio stems of the sound classes *truck* (c), *dogs* (d), *rain* (e), and *drilling* (f), as well as for the mixture (a). The colored plot (b) illustrates how sounds overlap in the time-frequency space.

cating them. As an additional self-imposed constraint of this study, the authors restrict themselves to single-channel audio recordings, suppressing the spatial perception of sound sources based on multi-channel information. Such single-microphone setup can be relevant for low-cost acoustic sensors.

As an introductory example, Fig. 1 illustrates a 5-s single-channel soundscape recording as Mel spectrogram [Fig. 1(a)], which includes sound events from the four sound classes "truck" engine, "dog" barking, "rain" drops, and "drilling" [Figs. 1(c)–1(f)]. Although each sound class shows distinctive spectral patterns, these patterns overlap as shown in the colored spectrogram plot [Fig. 1(b)]. A polyphony degree of four is assigned to this soundscape even though some of the sound classes appear as multiple sound events.

As the main contributions of this study, the human and machine performance for SPE are compared based on single-channel audio recordings. The results of a listening test are first presented and the human ability to count sound classes in polyphonic soundscapes is analyzed in detail. The authors then investigate the performance of several deep neural networks, which are trained to model SPE either explicitly or implicitly by first classifying active sound classes before counting them to compute the sound polyphony. The studied machine listening methods are based on Mel spectrogram input and include an adapted object detection algorithm, which is fine-tuned to recognize sound classes based on characteristic spectral patterns.

1 RELATED WORK ON AUDITORY COUNTING TASKS

1.1 Human Performance

Early psychoacoustic research involved listening tests based on simple stimuli, for example sinusoids and noise-like signals [7]. Natural audio signals such as speech, music, or everyday sounds are more complex in their spectral composition [8–10]. Despite their widely differing spectral properties, these signals are processed by the human auditory system using the same organizational principles [11]. The auditory perception of complex acoustic environments is challenging due to the unpredictable number of sound sources. At the same time, the number of auditory streams that humans can process simultaneously is limited to a maximum of four sources according to Kawashima et al. [12]. As a potential reason, Weisser [13] argues that the human auditory system gradually loses information during the processing of audio signals on the way from the ear periphery, stream formation, attention, to the final cognition step in order to avoid cognitive overflow and to better focus on particular sound sources.

Two studies investigated how the ability to recognize and annotate sound events is influenced by the number of concurrent sounds and their spectral characteristics. Carthwright et al. [14] showed that the human performance in annotating sound events decreases with increasing sound polyphony. As the main reason, the authors found that with more complex sound mixtures, humans tend to miss more and more relevant sound events (decreasing recall measure), whereas the annotation quality remains stable (stable precision measure). Piczak [15] found that the human sound event annotation performance is lower for noisy and ambient sounds and higher for more distinct sound sources as well as human and animal sounds. Given these results, the authors hypothesize that annotators tend to underestimate the sound polyphony degree in complex soundscapes and that the polyphony annotation performance is affected by the spectral characteristics of the audible sounds.

With a focus on music signals, Schöffler et al. [16] conducted a listening test on the task of counting musical instruments in short audio excerpts. Similar to this study, the stimuli covered polyphony degrees between one and six, which corresponds to one up to six different instruments per recording. The results showed that annotators could only reliably count up to three instruments. Similarly, humans can accurately segregate and count up to three simultaneous musical voices [17]. The performance decreases if the corresponding instrument timbres are more similar to each other [18].

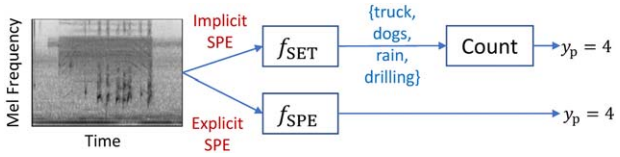

Fig. 2. SPE modeled explicitly using f_{SPE} (lower branch) or implicitly by counting the number of sound classes estimated by an SET model f_{SET} (upper branch).

1.2 Machine Performance

In this section, the authors will contrast the human performance on auditory counting tasks as discussed in the previous section with that of algorithms. Such algorithms often combine methods from audio signal processing and machine learning to estimate the number of sound sources.

In speech processing, the estimated number of speakers is often used to inform source separation algorithm in order to isolate each speaker signal. In contrast to natural soundscapes, the acoustic characteristics of different speakers can be very similar and therefore harder to distinguish. The number of speakers is commonly estimated explicitly using a multi-class classification approach [19, 20] based on deep neural networks or via clustering of sound sources in latent representation spaces [21–23].

The task of counting sound sources in soundscapes is commonly approached by analyzing multi-channel recordings. Several authors proposed methods for joint sound source counting, localization, and separation [24–26]. Because the spatial composition of a soundscape can be captured by the multi-channel recording setup, the number of (spatially distributed) sound sources can be estimated with high accuracy values of up to 89.8% for six sound sources [25]. Multimodal approaches, which analyze both audio and visual data by learning joint representation spaces, exist for tasks such as crowd counting [27, 28] and repetition counting [29].

2 TASK DEFINITION

In this study, the authors represent a *soundscape recording* using a single-channel audio sample vector $x \in \mathbb{R}^{N_s}$ of N_s samples. A soundscape commonly includes multiple *sound events*, each associated to one of N_c *sound classes*. A sound class activity vector is defined as $y_c \in \{0, 1\}^{N_c}$, which indicates whether at least one sound event in x is associated to the c-th sound class with $c \in [1, N_c]$. Further, the *sound polyphony degree* y_p is defined, and SPE is considered as the task of learning the mapping

$$f_{SPE} : x \in \mathbb{R}^{N_s} \mapsto y_p \in \mathbb{Z}. \quad (1)$$

As illustrated in Fig. 2, the authors distinguish between *explicit* SPE and *implicit* SPE (as in [30]). Whereas explicit SPE directly estimates y_p from x as in (1), implicit SPE first uses a (multi-label) SET model

$$f_{SET} : x \in \mathbb{R}^{N_s} \mapsto y_c \in \{0, 1\}^{N_c} \quad (2)$$

to estimate the sound class activity y_c for each of N_c sound classes based on the threshold sigmoid layer output of f_{SET}. Then, y_p is simply estimated from the number of active classes using

$$y_p = \sum_{i=1}^{N_c} (y_c)_i \, . \quad (3)$$

As confirmed by the majority of the listening test participants (compare SEC. 4), human listeners tend to approach the SPE task in an implicit fashion, detecting sounds prior to counting them.

3 DATASET

A manual annotation of the temporal boundaries of sound events in polyphonic soundscapes is labor-intensive and often further complicated by the overlap of concurrent sound events. As a consequence, many academic datasets include artificially generated soundscape recordings, which can be created in abundance by randomly mixing multiple isolated sound recordings [6, 31, 20]. In this study, the publicly-available Urban Sound Monitoring (USM) dataset [32],[1] which includes 24,000 polyphonic soundscapes of 5 s duration, was used. Every soundscape is created by mixing between one and six isolated sound recordings (*stems*) using random loudness ratios and spatial positions in a two-channel setup. A distinction is made between foreground and background sounds. Prior to the mixing, all stems are normalized to a perceived loudness of –12 dB LUFS based on ITU-R BS.1770-4 specification (see [32], SEC. 2.3 for details). The dataset covers 26 sound classes for urban sound monitoring, including vehicle sounds, construction site sounds, human-made sounds, animal sounds, climate sounds, and rare sounds such as sirens, gunshots, and church bells.

In these experiments, both stems ($y_p = 1$) and artificially mixed soundscapes ($y_p \in [2, 6]$) are combined, and a sound polyphony range between $y_p \in [1, 6]$ is considered. The pre-defined training/validation/test split of the USM dataset is adopted. As the only exception, a subset of the (monophonic) stems for $y_p = 1$ is randomly sampled from the training and validation sets to keep the polyphony classes approximately balanced during training. For the final model evaluation, all soundscapes of the evaluation set are considered.

Two issues, which complicate the SPE task, need to be discussed. First, the USM soundscapes are composed of foreground sounds and background sounds whose loudness levels were randomly sampled from different value ranges (compare SEC. 2.3, [32]). Some background sounds, which have a mixing coefficient close to the lower limit of −20 dB, are probably difficult to hear. Furthermore, the USM dataset builds upon audio samples uploaded to the collaborative audio database FreeSound.[2] Some of the

[1] https://github.com/jakobabesser/USM.
[2] https://freesound.org/.

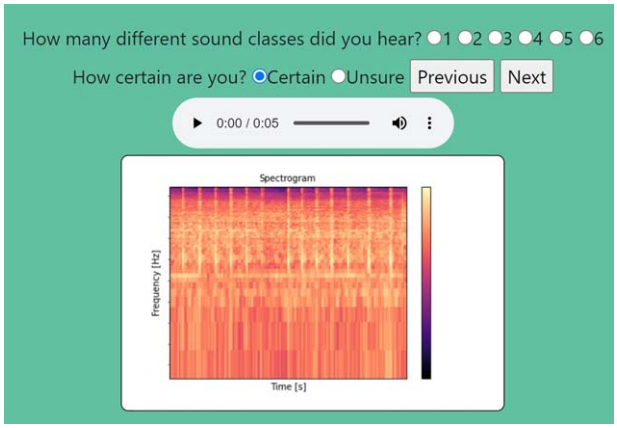

Fig. 3. Section of the graphical user interface of the listening test showing the Mel spectrogram of a test item. Participants are also presented with a list of all 26 sound classes of the USM dataset (not shown here).

audio examples exhibit unlabeled background noises in addition to the labeled sound events [33]. Such incomplete annotations can bias the human SPE performance because they increase the sound polyphony degree.

4 HUMAN PERFORMANCE

4.1 Listening Test

The authors conducted a listening test between November 2022 and January 2023 in a controlled lab environment to assess the human performance in SPE. Participants used closed headphones during the annotation and were guided through the listening test in a graphical interface displayed in a common web browser as illustrated in Fig. 3. The listening test was implemented in a locally hosted website using Django, HTML, CSS, and JavaScript. The website was optimized and tested for different web browsers on a desktop computer.

4.2 Participants & Stimuli

In total, 24 participants took part in this experiment, all of whom indicated prior experience from other listening tests. The number of participants per age group are 3 (18–24 years), 12 (25–34 years), 6 (35–44 years), 2 (45–54 years), and 1 (older than 55 years). The participants included part of the scientific staff at Fraunhofer Institute for Digital Media Technology as well as students at the Technische Universität in Ilmenau, Germany. No material incentive was given to participants.

All audio clips were randomly selected from the USM evaluation set taking into account a balanced polyphony degree (compare SEC. 3). The two-channel soundscapes were down-mixed by averaging both audio channels in order to remove any spatial cues for sound localization. A subset of the listening test stimuli can be accessed on an accompanying website [34]. All participants used closed headphones (Beyerdynamic DT 770) and could adjust the volume at any time during the listening test.

4.3 Training Procedure

Before starting the experiment, each participant underwent two voluntary training steps to familiarize themselves with the audio material and the annotation task. In the first step, participants could listen to 11 random sound stems from each of the 26 sound classes of the USM dataset. During playback, time-aligned visualizations of the audio signal's waveform and Mel spectrogram were provided.

This also allowed participants to learn a visual association between different sound classes and their temporal-spectral characteristics. In the second step, 11 randomly chosen soundscapes were provided as examples each for the polyphony degrees within $y_\mathrm{p} \in [2, 6]$ in order to get familiar with the main annotation task. All sound examples used in the training phase were taken from the training set of the USM dataset.

4.4 Test Procedure

For the listening test, the authors randomly selected from the USM test a set of 90 test items, which included 15 random isolated stems (polyphony degree of 1) and 15 random soundscapes from each sound polyphony degree within $y_\mathrm{p} \in [2, 6]$. From this collection, 15 test items were randomly assigned to each participant while ensuring a balanced distribution of polyphony degrees across all participants. In addition to the sound polyphony degree, the participants could annotate for each test item whether they felt certain or uncertain with their annotation. Furthermore, the authors constantly provided a list of all sound class labels during the annotation as reference. In total, each test item was annotated four times.

Inspired by [14], the authors considered three types of visual representations that were shown during the annotation. During the first group of five soundscapes, no additional visualization was shown. In the second group of five soundscapes, a waveform plot of the soundscape was shown. Finally, for the last group, a Mel spectrogram was shown.

4.5 Results

4.5.1 Influence of Annotation Certainty and Demographic Parameters

The authors first investigate the dependency between the SPE performance of the listening test participants and the certainty of their annotations. Fig. 4 shows in the left column the confusion matrices for SPE for all annotations [Fig. 4(a)] as well as separated between certain [Fig. 4(b)] and uncertain [Fig. 4(c)] annotations. The right column shows histograms over the estimation error $\epsilon_\mathrm{p} = \hat{y}_\mathrm{p} - y_\mathrm{p}$ with $\hat{y}_\mathrm{p}$ denoting the estimated polyphony degree. It is first observed that only up to a polyphony degree of $y_\mathrm{p} = 3$, the plurality of the corresponding annotations was correct. This result is in line with previous studies on similar auditory counting tasks as discussed in SEC. 1.1. When considering all annotations [Fig. 4(a)], the polyphony degree is underestimated on average by $\overline{\epsilon_\mathrm{p}} = -0.73$ (vertical line). As a contrary trend, it was found that in 31.5% of the uncertain annota-

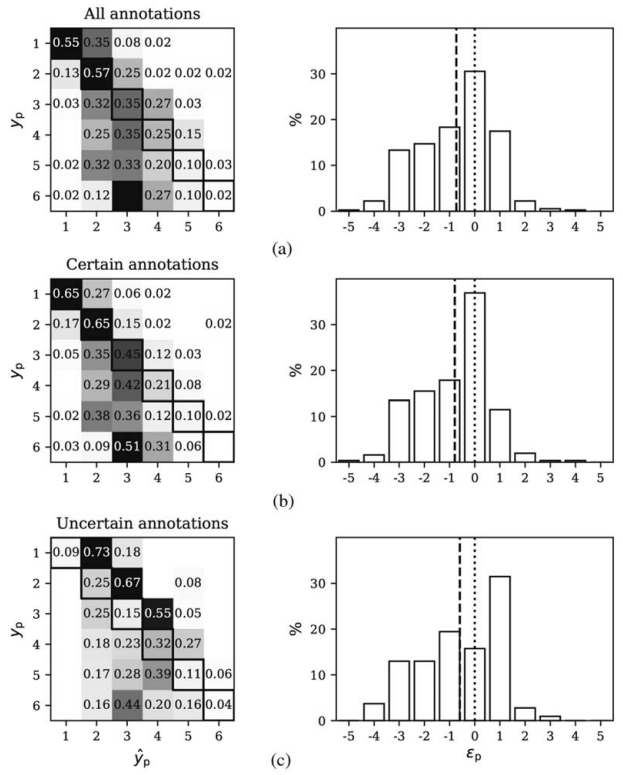

Fig. 4. Confusion matrices (left column) and histograms over the polyphony degree prediction error ε_P (right column) of listening test participants for all annotations (a) and separated by certain (b) and uncertain (c) annotations. Thicker bounding boxes indicate the percentage of correct estimations per polyphony degree (main diagonal). In the histogram plots, the dashed lines show the average over ε_P, and the dotted line indicates $\varepsilon_P = 0$.

tions [Fig. 4(c)], the sound polyphony was overestimated by $\varepsilon_p = 1$.

When looking into other annotation parameters and user demographics, the authors found no significant correlations between the SPE performance with the age group of the participants (Pearson correlation coefficient of $\rho = -0.31$, $p > 0.05$) or the training time prior to the listening test ($\rho = -0.08; p > 0.05$). The type of soundscape visualization (no visualization, waveform, or Mel spectrogram) did not have a significant influence on the participants' accuracy A_{SPE} as confirmed by a one-way ANOVA [$F(2, 69) = 0.05$; $p = 0.95$] and by the direct feedback of the participants. These results stand in contrast to the findings from [14], in which annotators perform better in SED when soundscapes were visualized as spectrograms. As mentioned in SEC. 4.2, the listening test participants were asked only whether they had participated in listening tests in the past, but not to what extent they were already familiar with spectrograms. Participants confirmed that particularly noisy sounds, such as wind or car, are difficult to detect and count visually based on their spectrograms.

4.5.2 Influence of Sound Characteristics

In this section, the authors aim to investigate which types of sounds are harder to count in polyphonic soundscapes than others. Naturally, the sound polyphony annotations only relate to the number but not the type of audible sound classes. As a consequence, the relationship between SPE performance and the presence and absence of particular sounds classes can only be investigated via an indirect approach. In the following, $N_f \in \mathbb{N}^+$ denotes the total number of test files used in the listening test, $N_a \in \mathbb{N}^{N_f}$ denotes the number of annotations per file, $\left(\hat{y}_{p,f}\right)_f \in \mathbb{N}^{(N_a)_f}$ for $f \in [1, N_f]$ denotes the sound polyphony annotations of file f, $y_{p,f} \in \mathbb{N}^{N_f}$ denotes the true sound polyphony per test file, and $y_{c,f} \in \{0, 1\}^{N_f \times N_c}$ denotes the sound class activity per file and sound class.

First, all annotations per file are averaged over to compute a file-level (absolute) polyphony estimation error $\varepsilon_{p,f} \in \mathbb{R}^{N_f}$ as

$$\left(\varepsilon_{p,f}\right)_f = \frac{1}{(N_a)_f} \sum_{a=1}^{(N_a)_f} \left|(\hat{y}_{p,f})_{f,a} - (y_{p,f})_f\right|. \tag{4}$$

Then, a class-level polyphony error $\varepsilon_{p,c} \in \mathbb{R}^{N_c}$ is computed as

$$\left(\varepsilon_{p,c}\right)_c = \frac{1}{N_c} \sum_{f=1}^{N_f} \left(\varepsilon_{p,f}\right)_f \cdot \left(y_{c,f}\right)_{f,c} \tag{5}$$

for $c \in [1, N_c]$.

Based on manual inspections of several sound examples, the 26 sound classes of the USM dataset are categorized into harmonic (H), transient (T), and noise-like (N) sounds as shown in Fig. 5. Then, the authors study how the number of active sound classes per category affect the file-level polyphony annotation error $\varepsilon_{p,f}$ for a given soundscape.

As a general trend, $\varepsilon_{p,f}$ increases with increasing polyphony degree y_p (Pearson correlation coefficient of $\rho = 0.78$; $p < 0.001$), which confirms that SPE becomes harder for soundscapes with higher polyphony. Interestingly, the number of noisy sound classes ($\rho = 0.62$; $p < 0.001$) followed by the number of transient sound classes ($\rho = 0.42$; $p < 0.001$) have the largest influence on the SPE performance. The authors conclude that because to their broadband spectral characteristics, sounds from these two sound class categories are harder to distinguish once they overlap. In contrast, the number of harmonic sound classes only shows a low correlation of $\rho = 0.24$ ($p < 0.01$) with $\varepsilon_{p,f}$. The authors conclude that concurrent harmonic sound events show a smaller amount of overlap because of their sparse energy distribution along frequency.

When looking at the class-level polyphony error $\varepsilon_{p,c}$ shown in Fig. 5, it is observed that salient but rare sounds such as screams, church bells, sirens, and gunshots coincide with better polyphony estimation, whereas stationary sounds such as construction site sounds (e.g., drilling or jackhammer) correlate with worse performance. Surprisingly, although most familiar to human listeners, the authors observe high error values for speech and cheering. Nevertheless, it was found that, contrary to the findings of [15] for SED, sparse sounds (H & T) are not easier to count than noise-like sounds (N), which is confirmed by a one-way ANOVA with no statistically significant differ-

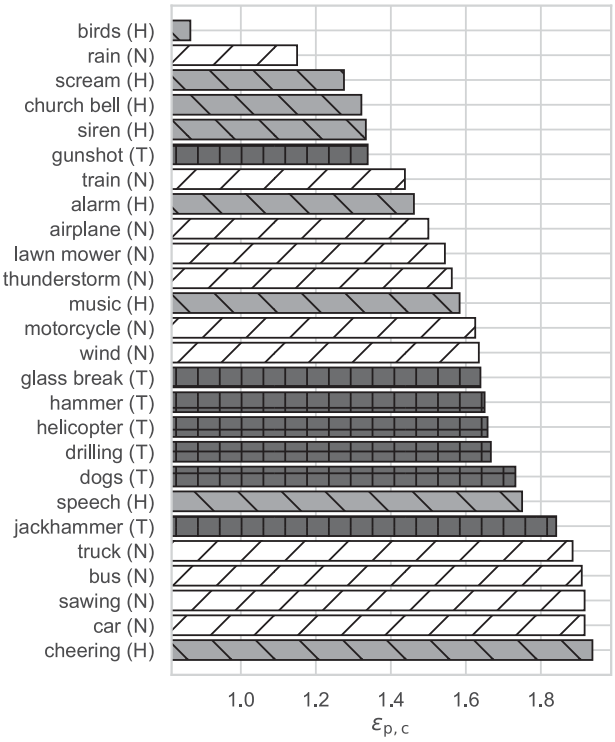

Fig. 5. Class-level polyphony error $\varepsilon_{p,c}$ sorted in ascending order. Class names are extended by a categorization of their spectral pattern as harmonic (H), transient (T), or noisy (N). Lower values indicate better performance.

ence in $\varepsilon_{p,c}$ between both groups of sound class categories [$F(1, 24) = 1.06$; $p = 0.31$]. Surprisingly, the smallest annotation error was observed for bird calls, which are usually short and prominent in the mid to high frequency spectrum.

5 MACHINE PERFORMANCE

5.1 Experimental Procedure

In these experiments, three deep neural network architectures for SPE and SET are evaluated. The architectures range from a small convolutional neural network (CNN) with 217,000 parameters [33], to the EfficientNetB0 architecture with 4.1 million parameters [35], to the YOLOv7 model with 37.3 million parameters [36].

The first two CNN variants described in SEC. 5.3 are trained for both explicit SPE and implicit SPE (compare SEC. 2). In the case of explicit SPE, each model is trained to predict the probability distribution $\hat{y}_p \in \mathbb{R}^6$ with $\hat{y}_p \in [0, 1]$ over all six sound polyphony classes using the categorical cross-entropy loss function. In the case of implicit SPE, the models are first trained for SET, i.e., to predict the individual probabilities $\hat{y}_c \in \mathbb{R}^{26}$ with $\hat{y}_c \in [0, 1]$ for all 26 sound classes using the binary cross-entropy loss function. Based on the SET results, the sound polyphony degree y_p is obtained by counting the detected sound classes with a probability of $\hat{y}_c > 0.5$. The training procedure used for the YOLOv7 models is detailed in SEC. 5.4. Here, implicit SPE is evaluated only based on the predicted bounding boxes for a given Mel spectrogram.

5.2 Feature Extraction

Throughout these experiments, the authors represent monaural audio recordings sampled at 22.05 kHz with Mel spectrograms using a hop size of 441 (20 ms), an FFT size of 1,024 (46.4 ms), and 128 Mel bands. Logarithmic magnitude scaling is applied to reduce the overall dynamic range between salient foreground and subtle background sounds. As a final step, the Mel spectrograms is normalized to the range [0, 1].

5.3 Convolutional Network Variants

The "Visual Geometry Group (VGG)-like" (`VGG`) model was used for SED in [33]. It contains six convolutional layers with an increasing number of filters from 32 to 128, a kernel size of 3×3, and intermediate max pooling operations. Global mean and max pooling are applied to the final feature map and concatenated as input to the final two dense layers. The model size is 217,000 parameters.

The authors further included the Pretrained Audio Neural Network (PANN) embeddings [37] because they have achieved state-of-the-art performance for sound event tagging on the AudioSet dataset [38]. They used the pre-trained "CNN14" model, which is composed of 12 convolutional layers and two dense layers. The 512-dimensional embedding vectors are used as input for a simple two-layer multilayer perceptron (`MLP`) model, which includes a dense layer with 128 neurons, a rectified linear unit activation function, and a final dense layer for classification. The MLP model has around 66,000 trainable parameters.

The EfficientNet architectures [35] were derived by simultaneously scaling the width, depth, and resolution of existing CNN architectures using a fixed convolution coefficient as a power to three constant values, while satisfying network complexity constraints. In these experiments, the authors test the EfficientNetB0 architecture, which combines regular convolutional layers with seven mobile inverted bottleneck layers that combine depth-wise separable convolutions and residual connections [39]. Two model variants are compared: The first variant (`EffNet`) is trained with random initial weights. The second variant (`EffNetIN`) is first pre-trained on the ImageNet [40] dataset for visual object recognition before all model layers are fine-tuned using the USM dataset.

Both CNN architectures were trained over 200 epochs using the Adam optimizer with a learning rate of 0.001. During training, the authors apply grid distortion (as implemented in the Albumentations Python library [41]) and SpecAugment [42] as data augmentation techniques to increase the variability of the training data. Grid distortion applies random spatial distortion locally in different areas of an image, and SpecAugment combines time and frequency masking. Both techniques are applied randomly with a probability of 0.5.

5.4 An Object Detection Approach

In computer vision, object detection algorithms localize objects by estimating a surrounding rectangular bounding box and the object category. In the audio domain, sound events exhibit characteristic patterns in time-frequency representations such as Mel spectrograms [10]. Unlike objects in natural images, these patterns are not necessarily invariant to operations such as scaling and shifting along frequency [43]. Although natural objects typically have closed contours, sound events have very different temporal-spectral shapes, ranging from sparse distributions of harmonic or transient sounds to texture-like distributions of ambient noise without well-defined temporal or frequency boundaries. As a result, the concept of bounding boxes for sound events in spectrogram representations is often ill-defined and ambiguous.

Nevertheless, in addition to the convolutional network variants described in SEC. 5.3, the YOLOv7 object detection algorithm [36] is used to detect sound events in Mel spectrograms based on characteristic spectral patterns. The network architecture includes a combination of standard convolutional layers, depth-wise convolutional layers, and skip connections. Additionally, YOLOv7 uses a multi-scale approach in which images are processed at multiple scales to improve detection accuracy. The network is trained using a combination of the mean square error loss, which penalizes bounding box coordinate estimation errors, and a loss term based on the intersection over union metric, which penalizes the difference between the predicted and ground truth class labels. This combination of losses measures the localization and classification performance simultaneously. One key aspect of YOLOv7 is the use of anchor boxes with pre-defined aspect ratios that allow the network to detect objects of various sizes.

5.4.1 Automatic Bounding Box Estimation

As discussed in SEC. 3, the USM dataset provides only file-level sound class annotations without temporal boundaries of sound events [32]. Because a manual bounding box annotation is too time-consuming, an automatic approach is used to estimate bounding boxes for training the YOLO network. The USM dataset provides the underlying sound class stems for each soundscape. Bounding boxes are therefore estimated from each individual stem, and all stem-level bounding boxes are combined as ground truth annotation for the soundscape. Each bounding box is described by the sound event onset time t_0 and offset time t_1, the lower and upper frequency boundaries f_0 and f_1, and the corresponding sound class y_c. The authors apply several methods provided by the OpenCV library [45].

First, a log-magnitude Mel spectrogram (see SEC. 5.2) is computed, and it was normalized to a range of [0, 255]. The resolution of the obtained "spectrogram image" is also upscaled from 128×216 to 640×640 using bicubic interpolation. A Gaussian blur low-pass filter with a kernel size of 7×7 is used to remove smaller artefacts.

Next, the spectrogram image is binarized to find contours that characterize the specific shape of sound event patterns in the Mel spectrogram. After initial experimentation, the authors decided to use a static binarization threshold of 127. The `cv2.findContours` method was then used to obtain a set of potential bounding box candidates from the binarized image. Finally, this set is refined, and all bounding box candidates whose area is less than the mean area of all bounding box candidates are removed. It has been found empirically that this dynamic threshold leads to good results, because it removes many erroneous small bounding box candidates and also adapts to each individual spectrogram.

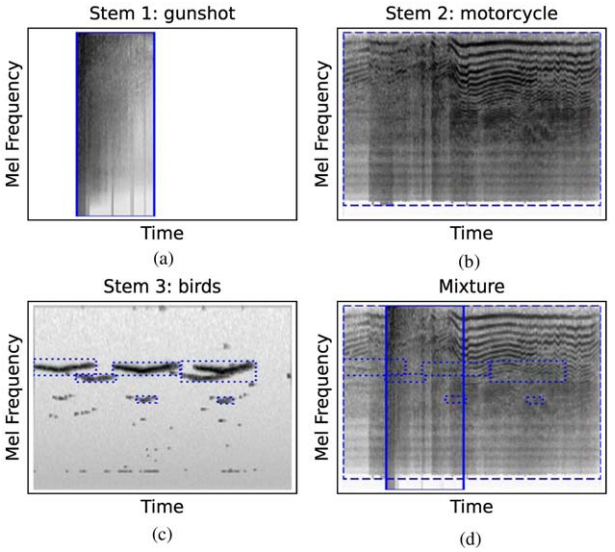

Fig. 6. Example results of the automatic bounding box estimation for a soundscape with a polyphony degree of $y_p = 3$ (USM training set, `10025.wav`). (a) Log-magnitude scaled Mel spectrograms are shown for the three underlying audio stems of the sound classes gunshot (solid lines), (b) motorcycle (dashed lines), and (c) birds (dotted lines), along with (d) the mixture.

This automatic labelling procedure naturally introduces label noise, which cannot be quantified at this stage. Around 50 training set files were manually inspected, and the estimated bounding boxes were found to be more reliable for sound classes with localized sounds such as dog barking, gunshots, and sirens, whereas the bounding boxes for ambient and noise-like sounds often spanned the entire time and frequency range.

This observation is illustrated in Fig. 6 for a training set example with a sound polyphony degree of $y_p = 3$. The first three subplots [Figs. 6(a)–6(c)] show the estimated bounding boxes for the three stems associated with the sound classes gunshot, motorcycle, and birds, and the fourth plot [Fig. 6(d)] shows the combination of all bounding boxes as annotation for the soundscape. The estimated bounding boxes for the noise-like motorcycle (dashed lines) and gunshot (solid line) sounds do not capture particular patterns but instead cover the full frequency and time extent. In contrast, the bounding boxes estimated from the bird recording well capture individual occurrences of the recurring bird call pattern (dotted lines).

Table 1. Method overview for SPE including the applied deep neural network architecture (second column), pre-training dataset (third column), and number of parameters (fourth column). Results are shown for implicit and explicit SPE based on the SET metric mAP_{SET} as well as the accuracy scores $A_{\text{SPE}}^{\text{Impl}}$ and $A_{\text{SPE}}^{\text{Expl}}$. $^{\uparrow}$Upper bound $A_{\text{SPE}}^{\text{Impl}}$ value using optimal class-wise thresholds is shown for best-performing model `EffNetIN`. Best results are shown in bold print.

Label	Network Architecture	Pre-Trained On	Trainable Parameters	Implicit SPE	Explicit SPE	
				mAP_{SET}	$A_{\text{SPE}}^{\text{Impl}}$	$A_{\text{SPE}}^{\text{Expl}}$
`VGG`	VGG-like CNN [33]	···	217,000	0.38	0.10	0.36
`PANN`	PANN [37] + MLP	AudioSet [38]	66,000	0.41	0.16	0.31
`EffNet`	EfficientNetB0 [35]	···	4.1 M	0.35	0.17	0.36
`EffNetIN`	EfficientNetB0 [35]	ImageNet [40]	4.1 M	**0.44**	0.19 (0.27$^{\uparrow}$)	**0.41**
`YoloTF`	YOLOv7 [36]	COCO [45]	37.3 M	0.29	**0.30**	···
`YoloT`	YOLOv7 [36]	COCO [45]	37.3 M	0.23	0.15	···
`Humans`	···	···	···	···	**0.31**	···

M = million.

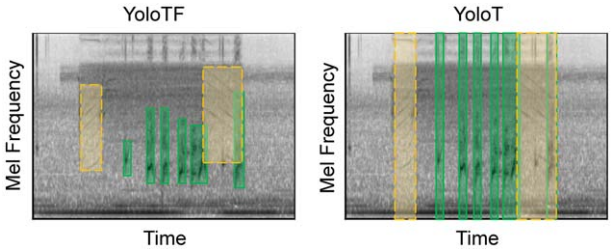

Fig. 7. Two approaches for bounding box estimation based on time and frequency boundaries (`YoloTF`, left) and only temporal boundaries (`YoloT`, right). For illustration purpose, the colored rectangles indicate a subset of all sound events in the polyphonic example shown in Fig. 1 including the sound classes dog barking (dashed line) and drilling (solid line).

5.4.2 Bounding Box Strategies

In this study, the YOLOv7 network is adapted for implicit SPE and first predict bounding boxes for individual sound events in a given Mel spectrogram and then estimate the sound polyphony degree y_{p} from the number of unique sound classes. Unlike bounding box estimation, SED algorithms only localize sound events in time. Therefore, two bounding box strategies as shown in Fig. 7 are compared: The `YoloTF` configuration considers the bounding box estimation over time and frequency, whereas the `YoloT` considers only the temporal localization of each sound event, similar to SED.

5.4.3 Network Training & Fine-Tuning

The YOLOv7 network used for the `YoloT` and `YoloTF` algorithms was pre-trained on the Common Objects in Context (COCO) dataset [45]. The COCO dataset is a widely-used benchmark dataset for object detection, segmentation, and image captioning. It contains over 330,000 images with 80 object categories and over 2.5 million object instances.

After its pre-training, the YOLO-based networks were fine-tuned on the USM training dataset based on Mel spectrogram "images" combined with bounding boxes estimated using the procedure described in SEC. 5.4.1. The networks were trained over 100 epochs with a batch size of eight images per batch. As the only difference, the frequency limits of all the bounding boxes used to train the `YoloT` algorithm were extended to the full frequency range of the Mel spectrogram.

6 RESULTS

Table 1 summarizes the performance of all methods for implicit and explicit SPE. As evaluation metrics, the SET performance is reported using the mean average precision mAP mAP_{SET} over all 26 sound classes in the USM dataset, and the SPE performance using the balanced accuracy scores $A_{\text{SPE}}^{\text{Expl}}$ and $A_{\text{SPE}}^{\text{Impl}}$ for explicit and implicit SPE, respectively.

The human SPE performance is provided as reference in the last row. The human performance is categorized as implicit SPE because all participants confirmed that they first tried to recognize sounds internally before counting them. Overall, the results demonstrate that SPE is very challenging for both humans and algorithms. The best-performing algorithms are the `EffNetIN` model, achieving $A = 0.41$ for explicit SPE, and the `YoloTF`, achieving $A = 0.30$ model for implicit SPE.

In general, explicit SPE seems to be the more promising strategy compared to implicit SPE. Although the compared deep neural network architectures perform very well in pattern recognition tasks such as object detection in natural images, their performance in SET is still limited, as confirmed by the best mean average precision of $mAP = 0.44$ achieved by the `EffNetIN` model. It is suspected that the number of remaining classification errors at this performance level is still too high to implicitly estimate sound polyphony based on the sound class predictions.

Fig. 8 contrasts the SPE confusion matrices obtained from the listening test and the best-performing `EffNetIN` model. It was observed that human listeners can estimate sound polyphony only up to a degree of $y_{\text{p}} = 3$ and perform poorly for higher polyphony degrees. In contrast, the `EffNetIN` model can distinguish better between soundscapes of higher polyphony degrees, which naturally also

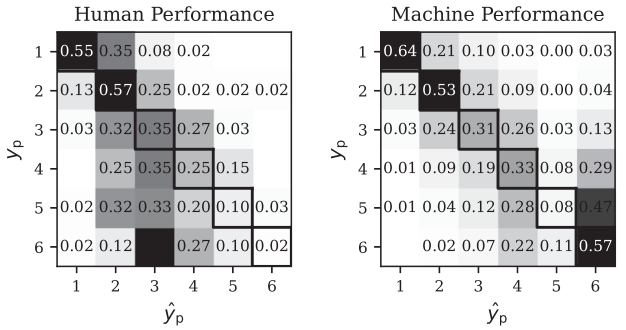

Fig. 8. SPE confusion matrices that contrast human performance with the best algorithm performance in explicit SPE obtained by the `EffNetIN` model (compare Table 1).

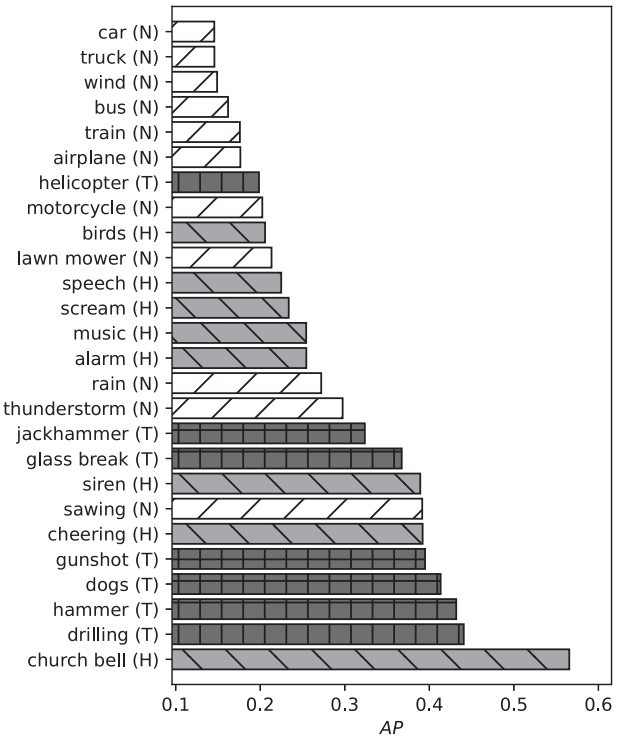

Fig. 9. Class-wise average precision scores for the `YoloTF` network for SET. Noise-like, transient, and harmonic sounds visualized using the same hatch encoding as before in Fig. 5. Higher values indicate better performance.

exhibit a stronger masking between different sound events. Most of the classifications errors happened between neighboring classes even though the relationship was not explicitly modeled in the loss function. This is in line with the results on speaker counting [19] and musical ensemble size classification [3].

In general, low accuracy scores of the CNN variants for implicit SPE were observed when using a fixed binarization threshold of 0.5 (see SEC. 5.1). To determine an upper bound, the authors examine the best-performing model `EffNetIN` and compute from the test set the optimal decision thresholds for each sound class that maximize the class-level f1 scores. These thresholds represent a hypothetical best-case scenario, because in common evaluation protocols, the test set cannot be used for hyperparameter optimization. By using these class-level thresholds, the $A_{\text{SPE}}^{\text{Impl}}$ metric improves by eight percent points from 0.19 to 0.27, which is still lower than the explicit SPE accuracy $A_{\text{SPE}}^{\text{Expl}} = 0.41$.

As shown by the `EffNetIN` model, pre-training on the ImageNet dataset is advantageous for SPE and SET, because the network can initially learn to recognize general two-dimensional patterns and then transfer and fine-tune this skill toward spectral patterns of different sound classes or polyphony levels. The `PANN` model is based on deep audio embeddings that are pre-trained on the AudioSet dataset. However, in contrast to the `EffNetIN` model, only the 66,000 parameters of the last MLP layers can be trained. Presumably, the significantly larger number of trainable parameters (4.1 million) allows the `EffNetIN` model to better fit the USM dataset.

Table 1 also shows that the CNN-based models outperform the YOLOv7-based models in SET. This supports the hypothesis that the bounding box concept is ill-defined for many sound classes, which is a key difference to objects in natural images. The annotation errors introduced by the automatic labeling described in SEC. 5.4.1 might be another reason for the performance difference. Comparing the `YoloTF` with the `YoloT` network, it is found that the former's ability to localize sounds not only in time but also in frequency leads to an improvement in both SET and SPE performance. When looking at the class-level SET performance of the `YoloTF` model shown in Fig. 9, it is observed that transient sounds such as drilling, hammer, dog barking, and gunshots on average are recognized best, followed by harmonic sounds such as sirens and speech, whereas it performs worst on noise-like sounds such as motorcycle and train. Particularly for noisy sounds, different tendencies were observed when comparing the human and machine performance on classifying and counting different sound classes. Most vehicle classes (car, truck, bus, train, airplane) and wind mostly have noise-like spectra, which are practically impossible to detect for the `YoloTF` model. Showing an average precision value of around 0.4, sawing seems to be an exception, because it often includes repetitive sound components. When comparing these tendencies to the human class-level SPE performance in Fig. 5, it is assumed that mainly the familiarity of listeners with particular sounds affects their ability to count them within mixtures of multiple sounds.

7 CONCLUSION

In this paper, the authors studied the task of estimating the sound polyphony of complex soundscapes, which were defined as the number of audible sound classes. They first reviewed scientific studies on both the human and machine performance in related auditory counting tasks such as instrument counting and speaker counting. In order to assess the human SPE performance, the authors carried out a listening test and studied in particular how the spectral characteristics of sound classes affect the human ability to count

them. In turn, they evaluated deep neural network architectures of different complexities for the SPE task, which were approached both explicitly by directly estimating the sound polyphony and implicitly by first classifying the presence of each type of sound before counting them. As an alternative approach for sound classification, a state-of-the-art computer vision algorithm was adapted for detecting objects in images.

Human listeners were able to reliably count up to three concurrent sound classes with the estimation error further increasing with increasing sound polyphony. In particular for soundscapes with a polyphony degree of up to three, listeners tended to underestimate the sound polyphony for annotations judged to be certain and overestimated the sound polyphony for uncertain annotations. Neither the age group, duration of the training phase, nor the type of audio visualization showed a significant influence on the SPE performance. Looking at the characteristics of the sounds, it was found that salient but infrequent sounds are the easiest to count, presumably because they attract more attention. The counting task becomes particularly challenging when several noise-like and transient sound classes overlap.

By combining pre-training using the ImageNet dataset with explicit SPE, the `EffNetIN` model could outperform the human listeners by 0.1 in accuracy and better distinguish between soundscapes of different sound polyphony degrees. Although implicit SPE does not seem to be feasible with the investigated neural network architectures, explicit SPE seems to be more promising. In general, these results confirm that sound classes with sparse spectral energy distribution over time (transient sounds) and frequency (harmonic sounds) are easier to classify by machine listening algorithms due to clearly pronounced spectral patterns.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the participants of the listening test and the authors of the NMF Toolbox [46] with the help of which the colored spectrogram plot in Fig. 1 was created. Furthermore, the authors would like to thank Andrew McLeod for constructive criticism of the manuscript. This study was supported by the German Research Foundation (AB 675/2-2) and has received funding by the European Union under the Horizon Europe vera.ai project, Grant Agreement number 101070093.

8 REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound* (The MIT Press, Cambridge, MA, 1990). https://doi.org/10.7551/mitpress/1486.001.0001.

[2] M. Kennedy, J. B. Kennedy, and T. Rutherford-Johnson (Eds), *The Oxford Dictionary of Music* (Oxford University Press, Oxford, UK, 2013), 6th ed. https://doi.org/10.1093/acref/9780199578108.001.0001.

[3] S. Grollmisch, E. Cano, F. Mora-Ángel, and G. López Gil, "Ensemble Size Classification in Colombian Andean String Music Recordings," in R. Kronland-Martinet, S. Ystad, and M. Aramaki (Eds.), *Perception, Representations, Image, Sound, Music*, pp. 60–74 (Springer, Cham, Switzerland, 2021).

[4] H. Cholakkal, G. Sun, F. S. Khan, and L. Shao, "Object Counting and Instance Segmentation With Image-Level Supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12397–12405 (Long Beach, CA) (2019 Jun.). https://doi.org/10.1109/CVPR.2019.01268.

[5] P. N. Amin, S. S. Moghe, S. N. Prabhakar, and C. M. Nehete, "Deep Learning Based Face Mask Detection and Crowd Counting," in *Proceedings of the 6th International Conference for Convergence in Technology (I2CT)*, pp. 1–5 (Maharashtra, India) (2021 Jan.). https://doi.org/10.1109/I2CT51068.2021.9417826.

[6] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A Library for Soundscape Synthesis and Augmentation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 344–348 (New Paltz, NY) (2017 Oct.). https://doi.org/10.1109/WASPAA.2017.8170052.

[7] D. Wang and G. L. Brown (Eds.), *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley-IEEE Press, Hoboken, NJ, 2006), 1st ed.

[8] B. Gold, N. Morgan, and D. Ellis (Eds.), *Speech and Audio Signal Processing: Processing and Perception of Speech and Music* (Wiley, Hoboken, NJ, 2011), 2nd ed. https://doi.org/10.1002/9781118142882.

[9] M. Müller, *Fundamentals of Music Processing Using Python and Jupyter Notebooks* (Springer, Cham, Switzerland, 2021), 2nd ed. https://doi.org/10.1007/978-3-030-69808-9.

[10] T. Virtanen, M. D. Plumbley, and D. E. Ellis, *Computational Analysis of Sound Scenes and Events* (Springer, Cham, Switzerland, 2018), 1st ed.

[11] M. Cooke and D. P. W. Ellis, "The Auditory Organization of Speech and Other Sources in Listeners and Computational Models," *Speech Commun.*, vol. 35, no. 3–4, pp. 141–177 (2001 Oct.). https://doi.org/10.1016/S0167-6393(00)00078-9.

[12] T. Kawashima and T. Sato, "Perceptual Limits in a Simulated 'Cocktail Party,'" *Atten. Percept. Psychophys.*, vol. 77, no. 6, pp. 2108–2120 (2015 Aug.). https://doi.org/10.3758/s13414-015-0910-9.

[13] A. Weisser, *Complex Acoustic Environments: Concepts, Methods and Auditory Perception*, Ph.D. thesis, Macquarie University, Sydney, Australia (2018 Sep.). https://doi.org/10.25949/19444259.v1.

[14] M. Cartwright, A. Seals, J. Salamon, et al., "Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations," *ACM Human-Comput. Interact.*, vol. 1, paper 29 (2017 Dec.). https://doi.org/10.1145/3134664.

[15] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018 (Brisbane, Australia) (2015 Oct.). https://doi.org/10.1145/2733373.2806390.

[16] M. Schoeffler, F.-R. Stöter, H. Bayerlein, B. Edler, and J. Herre, "An Experiment About Estimating the Number of Instruments in Polyphonic Music: A Comparison Between Internet and Laboratory Results." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 389–394 (Curitiba, Brazil) (2013 Nov.). https://doi.org/10.5281/zenodo.1417943.

[17] D. Huron, "Voice Denumerability in Polyphonic Music of Homogeneous Timbres," *Music Percept.*, vol. 6, no. 4, pp. 361–382 (1989 Jul.). https://doi.org/10.2307/40285438.

[18] P. Iverson, "Auditory Stream Segregation by Musical Timbre: Effects of Static and Dynamic Acoustic Attributes," *J. Exper. Psychol.: Hum. Percept.*, vol. 21, no. 4, pp. 751–763 (1995 Aug.). https://doi.org/10.1037//0096-1523.21.4.751.

[19] F.-R. Stöter, S. Chakrabarty, B. Edler, and E. A. Habets, "CountNet: Estimating the Number of Concurrent Speakers Using Supervised Learning," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 2, pp. 268–282 (2019 Feb.). https://doi.org/10.1109/TASLP.2018.2877892.

[20] M. Yousefi and J. H. Hansen, "Real-Time Speaker Counting in a Cocktail Party Scenario Using Attention-Guided Convolutional Neural Network," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1484–1488 (Brno, Czechia) (2021 Aug./Sep.).

[21] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35 (Shanghai, China) (2016 Mar.). https://doi.org/10.1109/ICASSP.2016.7471631.

[22] Z. Chen, Y. Luo, and N. Mesgarani, "Deep Attractor Network for Single-Microphone Speaker Separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 246–250 (New Orleans, LA) (2017 Mar.). https://doi.org/10.1109/ICASSP.2017.7952155.

[23] Y. Xiao and H. Zhang, "Improved Source Counting and Separation for Monaural Mixture," *arXiv preprint arXiv:2004.00175* (2020 Apr.).

[24] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-Time Multiple Sound Source Localization and Counting Using a Circular Microphone Array," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 10, pp. 2193–2206 (2013 Oct.). https://doi.org/10.1109/TASL.2013.2272524.

[25] M. Jia, J. Sun, and C. Bao, "Real-Time Multiple Sound Source Localization and Counting Using a Soundfield Microphone." *J. Ambient Intell. Humaniz. Comput.*, vol. 8, pp. 829–844 (2016 Jun.). https://doi.org/10.1007/s12652-016-0388-x.

[26] B. Yang, H. Liu, C. Pang, and X. Li, "Multiple Sound Source Counting and Localization Based on TF-Wise Spatial Spectrum Clustering," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1241–1255 (2019 Aug.). https://doi.org/10.1109/TASLP.2019.2915785.

[27] D. Hu, L. Mou, Q. Wang, et al., "Ambient Sound Helps: Audiovisual Crowd Counting in Extreme Conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2249–2259 (Online) (2021 Jun.).

[28] U. Sajid, X. Chen, H. Sajid, T. Kim, and G. Wang, "Audio-Visual Transformer Based Crowd Counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2249–2259 (Montreal, Canada) (2021 Oct.).

[29] Y. Zhang, L. Shao, and C. G. M. Snoek, "Repetitive Activity Counting by Sight and Sound," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14070–14079 (Nashville, TN) (2021 Jun.). https://doi.org/10.1109/CVPR46437.2021.01385.

[30] M. Taenzer, S. I. Mimilakis, and J. Abeßer, "Informing Piano Multi-Pitch Estimation With Inferred Local Polyphony Based on Convolutional Neural Networks," *Electronics*, vol. 10, no. 7, paper 851 (2021 Apr.). https://doi.org/10.3390/electronics10070851.

[31] G. Lafay, M. Rossignol, N. Misdariis, M. Lagrange, and J.-F. Petiot, "Investigating Soundscapes Perception Through Acoustic Scenes Simulation," *Behav. Res. Methods*, vol. 51, pp. 532–555 (2018 Oct.).

[32] J. Abeßer, "Classifying Sounds in Polyphonic Urban Sound Scenes," presented at the *152nd Convention of the Audio Engineering Society* (2022 May), paper 10570.

[33] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 829–852 (2022). https://doi.org/10.1109/TASLP.2021.3133208.

[34] J. Abeßer, A. Ullah, S. Ziegler, and S. Grollmisch, "Listening Test Stimuli (Online Demo)," https://machinelistening.github.io/JAES_2023_SPE_ListeningTestExamples.html (2023).

[35] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114 (Long Beach, CA) (2019 Jun.).

[36] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," *arXiv preprint arXiv:2207.02696* (2022 Jul.). https://doi.org/10.48550/arXiv.2207.02696.

[37] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2880–2894 (2021). https://doi.org/10.1109/TASLP.2020.3030497.

[38] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, et al., "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*

(ICASSP), pp. 776–780 (New Orleans, LA) (2017 Mar.). https://doi.org/10.1109/ICASSP.2017.7952261.

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520 (Salt Lake City, UT) (2018 Jun.). https://doi.org/10.1109/CVPR.2018.00474.

[40] J. Deng, W. Dong, R. Socher, et al., "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255 (Miami, FL) (2009 Jun.). https://doi.org/10.1109/CVPR.2009.5206848.

[41] A. Buslaev, A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, paper 125 (2020 Feb.). https://doi.org/10.3390/info11020125.

[42] D. S. Park, W. Chan, Y. Zhang, et al., "SpecAugment: A Simple Augmentation Method for Automatic Speech Recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2613–2617 (Graz, Austria) (2019 Sep.). https://doi.org/10.21437/Interspeech.2019-2680.

[43] P. Pham, J. Li, J. Szurley, and S. Das, "Eventness: Object Detection on Spectrograms for Temporal Localization of Audio Events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2491–2495 (Calgary, Canada) (2018 Apr.). https://doi.org/10.1109/ICASSP.2018.8462062.

[44] Itseez, "Open Source Computer Vision Library," https://github.com/itseez/opencv (2015).

[45] T.-Y. Lin, M. Maire, S. Belongie, et al., "Microsoft COCO: Common Objects in Context," in *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, pp. 740–755 (Zurich, Switzerland) (2014 Sep.).

[46] P. López-Serrano, C. Dittmar, Y. Özer, and M. Müller, "NMF Toolbox: Music Processing Applications of Nonnegative Matrix Factorization," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (Birmingham, UK) (2019 Sep.).

THE AUTHORS


Jakob Abeßer


Asad Ullah


Sebastian Ziegler


Sascha Grollmisch

Jakob Abeßer received the Diploma degree in computer engineering from the Technische Universität Ilmenau, Germany in 2008. Since 2008, he has been working as research scientist at the Fraunhofer Institute for Digital Media Technology (IDMT), Ilmenau, Germany. In 2010, he visited the Centre of Excellence in Music, Mind, Body and Brain, University of Jyäskylä, Finland, for a research stay. In 2014, he received his Ph.D. degree in media technology under supervision of Prof. Gerald Schuller. From 2012 to 2017, he further joined the University of Music Franz Liszt in Weimar, Germany, as a research associate in the Jazzomat research project under supervision of Prof. Martin Pfleiderer. Since 2018, he has been a principal investigator, and since 2021, he has been a senior scientist at Fraunhofer IDMT. He is currently also a visiting researcher in the Semantic Audio Processing Group headed by Prof. Meinard Müller at the International Audio Laboratories in Erlangen, Germany. Working toward a habilitation degree, his current research focuses on machine listening and music information retrieval.

•

Asad Ullah worked as a research assistant in the Semantic Music Technologies group at Fraunhofer Institute for Digital Media Technology (IDMT). He is pursuing his Master's in media technology at TU Ilmenau. He received his Bachelor's in mechatronics engineering from the University of Engineering and Technology in Peshawar, Pakistan, in 2016. He worked as a research associate at Data Analytics Lab, LUMS, Pakistan, under the supervision of Dr. Imdad Ullah Khan, where he worked on various research projects. Currently, he is working toward his Master's degree focusing on bio-acoustics sound event detection using few-shot learning. His research interests include deep learning, computer vision, biomedical signals, and medical image analysis.

•

Sebastian Ziegler is an undergraduate at the Technische Universität in Ilmenau. He is currently doing an internship at the Fraunhofer Institute for Digital Media Technology (IDMT) Ilmenau, Germany, and pursuing a Bachelor's degree in media technology.

•

After finishing his engineering diploma in Media Technology in 2009 at Technische Universität Ilmenau, Sascha Grollmisch started his career as a software developer at Fraunhofer Institute for Digital Media Technology (IDMT). He was later part of the spin-off company Songquito, which distributes the music education software Songs2See, developed within a long-term research project at Fraunhofer IDMT. For their effort in developing one of the first fully interactive music learning games, the Songquito team received the Innovation and Entrepreneur Award of the German Informatics Society. In the following years, Sascha's interest and knowledge in automatic music and audio analysis grew stronger with several industry projects, changing his role from software developer to deep learning researcher. With the ACMus research project, Sascha started his Ph.D. at Technische Universität Ilmenau in 2019. His thesis focuses on few-shot and semi-supervised deep learning for audio classification tasks from industrial sounds to music recordings.