How Robust are Audio Embeddings for Polyphonic Sound Event Tagging?

Jakob Abeßer, *Member, IEEE*, Sascha Grollmisch, and Meinard Müller, *Fellow, IEEE*

***Abstract*—Sound classification algorithms are challenged by the natural variability of everyday sounds, particularly for large sound class taxonomies. In order to be applicable in real-life environments, such algorithms must also be able to handle polyphonic scenarios, where simultaneously occurring and overlapping sound events need to be classified. With the rapid progress of deep learning, several deep audio embeddings (DAEs) have been proposed as pre-trained feature representations for sound classification. In this article, we analyze the embedding spaces of two non-trainable audio representations (NTARs) and five DAEs for sound classification in polyphonic scenarios (sound event tagging) and make several contributions. First, we compare general properties like the inter-correlation between feature dimensions and the scattering of sound classes in the embedding spaces. Second, we test the robustness of the embeddings against several audio degradations and propose two sensitivity measures based on a class-agnostic and a class-centric view on the resulting drift in the embedding space. Finally, as a central contribution, we study how a blending between pairs of sounds maps to embedding space trajectories and how the path of these trajectories can cause classification errors due to their proximity to other sound classes. Throughout our analyses, the PANN embeddings have shown the best overall performance for low-polyphony sound event tagging.**

***Index Terms*—Sound event tagging, sound polyphony, deep audio embeddings, embedding space.**

I. Introduction

THE ability to recognize sounds is of vital importance for navigating through different everyday environments. Each environment comes with its unique set of sounds whose detection and categorization is an essential part of auditory scene analysis. While sound event detection aims at localizing sound events in time, sound event tagging (SET) focuses solely on classifying all sound classes occurring in a given scene [1].

Sounds from the same class can exhibit large differences in timbre, duration, and loudness. This intrinsic variability already makes the classification of isolated sound events (sound event classification) a challenging task. Real-life environments are often characterized by multiple sound sources, which are audible at the same time. In this article, we focus on the challenge of classifying overlapping sounds in such scenarios, a task which we refer to as sound event tagging (SET).

Deep neural networks, which are a core component of state-of-the-art sound classification and tagging algorithms, require large amounts of training data if they are trained in a supervised fashion. In many application scenarios however, only limited amounts of annotated data are available. Transfer learning has been successfully used for SET [2], [3], [4], [5], [6] to pre-train deep neural networks on large datasets and later fine-tune them for novel (down-stream) tasks with limited amounts of training data. The intermediate layer representations of such networks (embeddings) have been shown to be powerful features for several audio classification tasks [7] and related tasks such as audio source separation [8] and acoustic scene classification [9].

As the main contribution of this article, we compare various (pre-trained) audio embeddings for SET, i. e., sound event classification in polyphonic scenarios. We focus our investigations on lower sound polyphony degrees and study how mixtures of different sound classes are represented in the embedding spaces of two non-trainable audio representations (NTARs) and five deep audio embeddings (DAEs), which are pre-trained and then applied for SET. Second, we test the robustness of the embeddings against three different types of audio degradations, which are common in real-life sound monitoring applications. To this end, we propose to measure the resulting embedding drift in the embedding space both from a class-agnostic and from a class-centric view. Finally, as a central contribution, we investigate how overlapping sounds are represented in the embedding spaces. For this we implement a continuous blending between sound pairs and study the resulting trajectories in the embedding space. We aim to understand how the path of these trajectories can cause sound misclassification due to its proximity to other sound classes. Fig. 1 illustrates how audio degradations (middle) and blending between sounds (bottom) may influence the audio clip's position in the embedding space. In the first example, a car sound is first modified to have a lower volume and then mixed with ambient background sounds. In the second example, an alarm sound is continuously blended with a car sound. As illustrated, the resulting shifts and trajectories in the embedding space can cause confusions with other sound classes (e. g., bird calls).

The remainder of this article is organized as follows. Section II provides an overview of the relevant scientific work. Section III

Manuscript received 2 November 2022; revised 15 May 2023 and 16 June 2023; accepted 28 June 2023. Date of publication 11 July 2023; date of current version 14 July 2023. This work was supported by the German Research Foundation (AB 675/2-2, MU 2686/11-2). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Emmanouil Benetos. *(Corresponding author: Jakob Abeßer.)*

Jakob Abeßer is with the Fraunhofer IDMT, 98693 Ilmenau, Germany, and also with the International Audio Laboratories Erlangen, 91058 Erlangen, Germany (e-mail: jakob.abesser@idmt.fraunhofer.de).

Sascha Grollmisch is with the Fraunhofer IDMT, 98693 Ilmenau, Germany (e-mail: sascha.grollmisch@idmt.fraunhofer.de).

Meinard Müller is with the International Audio Laboratories Erlangen, 91058 Erlangen, Germany (e-mail: meinard.mueller@audiolabs-erlangen.de).

Digital Object Identifier 10.1109/TASLP.2023.3293032

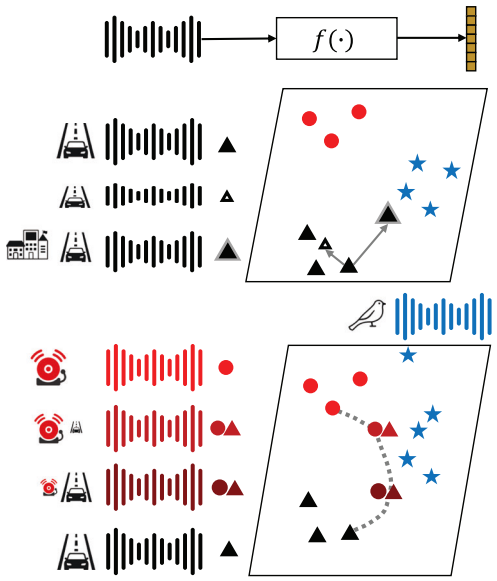

Fig. 1. Mapping of audio clips to embedding representations (top). Measuring the robustness of audio embeddings against audio degradations such as loudness variations and background noise (middle) and for mixtures of overlapping sound events (bottom).

describes the general procedure for extracting embeddings from audio signals. Furthermore, this section introduces the NTARs and DAEs compared in this article. Section IV explains how we generate and augment audio recordings with overlapping sounds to serve as a dataset. As the main part of this article, we discuss methods for exploring different embedding spaces (see Section V) and the robustness of embedding representations with respect to degradations of the audio signal (see Section VI). Furthermore, we study in Section VII the embedding space trajectories of blended sounds. Finally, Section VIII concludes this article.

We publish relevant data and source code alongside this article to enable reproducibility of the experiments.[1]

II. Related Work

The paradigm of transfer learning was successfully used in various disciplines ranging from computer vision, natural language processing, to speech processing [10]. In the field of audio analysis, DAEs were trained either in a supervised or self-supervised fashion [11]. A common self-supervised learning strategy is contrastive learning [12], [13], [14], where embedding representations are learnt to capture similarity relationships between data instances or augmented versions thereof. While most DAEs operate solely in the audio domain, relationships between audio data and other modalities can be modeled by learning joint embedding spaces. Such cross-modal embeddings were applied for several audio-visual tasks such as identity verification [15], audio-visual stream correspondence [4], scene classification [16], text-based audio retrieval [14], [17], [18], and cross-modal retrieval based on audio, images, and text [13]. Further knowledge and constraints can be integrated during the learning process, for instance, using additional loss terms for regularization [19].

While most deep audio embeddings rely on spectrogram-like feature representations such as the Mel-spectrogram [2], [3], [4], [12], Lopez-Meyer et al. [20] propose a convolutional neural network (CNN) architecture that maps raw audio clips to an embedding representation in an end-to-end fashion. Kong et al. [5] combine both waveform-based and spectrogram-based features in deriving the PANN embeddings. Deep generative models for audio synthesis [21] or music synthesis [22] on a waveform-level often use encoder–decoder network architectures to learn suitable embedding representations [23], which can further be be regularized to control the perceptual properties of the synthesized audio [24].

DAEs have been applied for a large variety of down-stream tasks. Most audio embeddings are trained on the AudioSet [25], which to date is the largest set of audio files from different domains, including speech, environmental sounds, and music. Previous work has shown that DAEs trained for sound classification perform well for related tasks such as urban sound tagging [26], [27], [28], acoustic scene classification [9], and various other audio classification tasks ranging from music to industrial sounds [7]. Notably, DAEs are also effective for tasks outside of their original training data domain such as audio captioning [29], [30], speech enhancement [31], and for detecting COVID-19 in respiratory-related sounds like breathing, cough, and speech [32]. Furthermore, DAEs have been used for SED in order to inform algorithms for source separation [8], [33] and speech denoising algorithms [34].

In the context of transfer learning, the most common way to evaluate embedding representations is to measure their performance on a set of down-stream tasks [4], [5], [7]. The Holistic Evaluation of Audio Representations (HEAR) benchmark represents the largest effort so far to evaluate embeddings for a large number of down-stream tasks [35].

In addition to such general performance evaluations, embedding spaces have been investigated to better understand the predictions of classification models. For multi-class classification tasks, it is common to visualize embedding spaces after applying dimensionality reduction techniques such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), or Uniform Manifold Approximation and Projection (UMAP). Such visualizations allow for testing whether class instances form well separated clusters in the embedding space. A common observation is that class separability typically improves in the embedding space from layer to layer, which supports the idea of hierarchical feature learning [36]. We study in detail how sound classes scatter in the embedding spaces of different audio representations in Section V-B.

The analysis of DAEs can provide powerful cues about characteristics of the input data of a neural network. As shown by Stacke et al. in [37], a discrepancy between embedding space distributions of two datasets can be used as a proxy to quantify domain shift. Similarly, changes in the embedding space have been investigated as indicator for the robustness of embedding

[1][Online]. Available: https://github.com/jakobabesser/embedding_robustness_2022

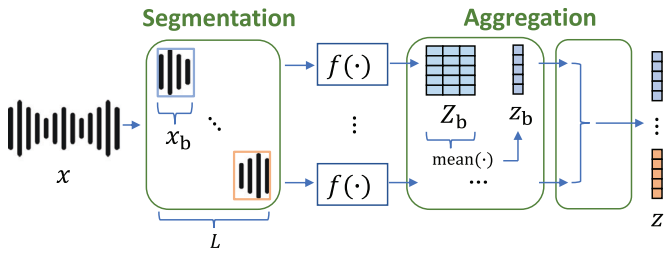

Fig. 2. Audio clips are segmented into one-second blocks from which embedding vectors are extracted and averaged. Block-level embedding vectors are then concatenated to yield a final embedding vector.

TABLE I
COMPARISON OF ALL COMPARED AUDIO REPRESENTATIONS IN TERMS OF THE BLOCK-LEVEL EMBEDDING MATRIX SIZE E_b AND FEATURE RATE M, THE STACKED EMBEDDING SIZE E, AS WELL AS THE TRAINING OBJECTIVE OF THE DAES (SL - SUPERVISED LEARNING, SSL - SELF-SUPERVISED LEARNING)

Type	E_b	M [Hz]	E	Training
`MelSpec`	128	22	640	-
`MFCC`	13	22	65	-
`Kumar` [3]	1024	1	5120	SL
`OpenL3` [4]	512	42	2560	SSL
`PANN` [5]	512	1	2560	SL
`PaSST` [6]	1295	20	6475	SL
`VGGish` [2]	128	1	640	SL

representations towards degradations of the audio signal [38]. While this previous study investigated only the OpenL3 and YamNet embeddings, we follow in this article a similar approach to measure the robustness of seven different audio representations towards audio degredations in Section VI. While most prior work focused on multi-class classification tasks, the study of embedding spaces for multi-label tasks such as SET is a relatively new field of research [39], [40]. To the best of our knowledge, no prior work investigated embedding spaces for SET.

III. AUDIO REPRESENTATIONS

In this section, we first explain the general procedure of embedding extraction from audio files. Then, we introduce seven audio representations, which we compare in our experiments. These representations include two non-trainable audio representations (NTARs) as discussed in Section III-B and five pre-trained deep audio embeddings (DAEs) as discussed in Section III-C.

A. Embedding Extraction

In the following, we explain how we extract embedding vectors from monaural audio clips $x \in \mathbb{R}^{L \cdot f_\mathrm{s}}$. We enforce the clip duration to be an integer multiple of seconds by applying zero-padding if necessary. $L \in \mathbb{N}$ denotes the number of seconds and $f_\mathrm{s} \in \mathbb{N}$ denotes the sample rate in Hz. As visualized in Fig. 2, we first partition the audio clip x into L non-overlapping blocks $x_\mathrm{b} \in \mathbb{R}^{f_\mathrm{s}}$ of one-second duration. An embedding function f maps each block x_b to a block-level embedding matrix Z_b as $f : x_\mathrm{b} \in \mathbb{R}^{f_\mathrm{s}} \rightarrow Z_\mathrm{b} \in \mathbb{R}^{E_\mathrm{b} \times M}$ with $E_\mathrm{b} \in \mathbb{N}$ denoting the embedding size and $M \in \mathbb{N}$ denoting the feature rate in Hz. Afterwards, we average Z_b over the time frames and obtain an embedding vector $z_\mathrm{b} \in \mathbb{R}^{E_\mathrm{b}}$. Finally, we stack all block-level embedding vectors to a final embedding vector $z \in \mathbb{R}^{E}$ with $E = L \cdot E_\mathrm{b}$. In our experiments, we analyze 5s long audio files, hence $L = 5$. As an alternative, variable-length input clips could be processed using a shingle-based approach [41], where multiple pre-defined fixed-size embedding matrices are extracted from longer audio clips using an overlap of 50 %.

When analyzing a set of $N \in \mathbb{N}$ audio clips, we stack their embedding vectors to an embedding matrix $Z \in \mathbb{R}^{N \times E}$. As basis for distance calculations in the corresponding embedding space, we apply z-score normalization to Z. As will be discussed in the following sections, most investigated audio representations have different time resolutions. The presented approach of averaging over block-level embeddings leads to the same temporal resolution of one second for each embedding, which we believe is a good compromise that allows to capture time-dependent sound characteristics.

Table I summarizes all audio representations, which are compared in this article: The embedding dimensionality E_b and feature rate M of the block-level embedding matrices Z_b as well as the embedding dimension E of the stacked embeddings z are provided. The training objective (last column) of the DAEs is either supervised learning (SL) or self-supervised learning (SSL).

B. Non-Trainable Audio Representations

As baseline representations, we use the *librosa* Python library [42] to compute two NTARs, which characterize the time-frequency energy distribution in the audio clips. Here, we use a sample rate of $f_\mathrm{s} = 22.05$ kHz (as in [43]), a hopsize of 1024 samples (46.4ms), and a blocksize of 2048 samples (92.9ms). The first representation is a log-magnitude Mel-spectrogram (`MelSpec`) using 128 Mel bands. As second representation, we compute the first 13 Mel-frequency Cepstral Coefficients (`MFCC`) as a compact representation of the spectral envelope. Both NTARs have a feature rate of $M = 22$ Hz.

C. Deep Audio Embeddings

In addition to the NTARs introduced in Section III-B, we investigate five pre-trained DAEs, which are based on different CNN and Transformer architectures. All DAEs were trained on the AudioSet dataset [25], which is the largest audio dataset to date covering different audio domains. The AudioSet includes around two million audio clips, which are weakly-labeled with an average of 2.7 labels per file. The dataset covers 527 sound classes. All DAEs except for the `PANN` embeddings use Mel-spectrogram variants as input features, however with different number of Mel bands and time resolution (hopsize). As shown in the original publications, the performance of the DAEs is greatly influenced by their parameters. We use the best-performing models here and do not include further ablation studies related to model parameters. While this section provides a high-level

overview over the applied deep audio embeddings, a detailed list of model parameters are provided on the accompanying website. Since all DAEs rely on NTARs as input representations, we hypothesize that DAEs in general will show a better performance on the SET task.

Kumar et al. [3] proposed a CNN with 12 convolutional layers with intermediate pooling and a final global pooling operation to make use of the weak labels of the AudioSet. The network processes Mel-spectrograms with 128 Mel bands as input representations. Finally, the layer activations of the penultimate convolutional layer are used as (`Kumar`) embedding vector with $E_\mathrm{b} = 1024$ and a feature rate of $M = 1$ Hz.

The `OpenL3` [4] embeddings are DAEs that are trained in a self-supervised fashion. This approach does not require any labeled data but instead uses audio–visual correspondences as training objective. The underlying L^3-Net was initially proposed in [44] and includes two sub-networks for audio and image processing, respectively, and several fusion layers. The audio sub-network processes Mel-spectrograms using a stack of four convolutional layers with intermediate max pooling. Multiple configurations of the `OpenL3` embeddings exist which were trained on different subsets of the AudioSet dataset. We use the "music" configuration with 256 Mel bands and an embedding size of $E_\mathrm{b} = 512$, which has shown to outperform the "environmental" configuration for various datasets including the ESC-50 dataset [4], [7]. The embeddings have a feature rate of $M = 42$ Hz.

The Pretrained Audio Neural Network (`PANN`) embeddings were introduced by Kong et al. [5]. Among several tested network architectures, the "Wavegram-Logmel-CNN" model performed best for the AudioSet (sound) tagging task. The used CNN14 model includes a total of 12 convolutional layers combined with two final dense layers. Opposed to the other three DAEs, the `PANN` embeddings combine as input a learnable waveform-based input feature (wavegram) and a non-trainable Mel-spectrogram with 64 Mel bands. Furthermore, a final global pooling operation aggregates the full temporal context of a given audio file by combining max and average pooling. The final dimensionality of the `PANN` embedding matrix is $E_\mathrm{b} = 512$ with a feature rate of $M = 1$ Hz.

The `VGGish` embeddings [2] are based on a modified version of a VGG model [45] that includes five convolutional layers and three final dense layers. The network takes log-magnitude Mel-spectrograms with 64 Mel bands as input. Each `VGGish` embedding vector has a size of $E_\mathrm{b} = 128$ with a feature rate of $M = 1$ Hz.

As alternative to the convolutional neural network architecture, we incorporate the Audio Spectrogram Transformer (AST) model [46] as DAE, which takes sequences of Mel-spectrogram patches as input. In particular, we use the PaSST-S model proposed in [6], which was trained using a strategy referred to as structured patchout. The patchout technique involves removing randomly chosen patches from the input sequence. In structured patchout, the removed patches are selected in such way that they cover the entire frequency range at one particular time window or, vice versa, the entire clip duration at a specific frequency range. This approach is comparable to data augmentation techniques used for SpecAugment [47]. Each `PaSST` embedding vector has a size of $E_\mathrm{b} = 1295$ with a feature rate of $M = 20$ Hz.

TABLE II
DESCRIPTIONS AND LABELS FOR FIVE DIFFERENT AUDIO DEGRADATIONS WITH CORRESPONDING *AUDIOMENTATIONS* PARAMETER SETTINGS BELOW

a	**Description (*Label*)**
1	No degradation
2	Loudness reduction (*Quiet*) `Gain(gain_in_db=-10)`
3	Gaussian Noise (*Noise*) `AddGaussianSNR(snr_in_db=10)`

4	High-Frequency Boost (*BoostHigh*)
	`LowShelfFilter(center_freq=100, gain_db=-10, q=0.3)`
	`HighShelfFilter(center_freq=2500, gain_db=10, q=0.3)`

5	Low-Frequency Boost (*BoostLow*)
	`LowShelfFilter(center_freq=100, gain_db=10, q=0.3)`
	`HighShelfFilter(center_freq=2500, gain_db=-10, q=0.3)`

IV. EXPERIMENTAL DATASET

For our investigations, we use an augmented version of the ESC-50 dataset [48], which is a freely-available sound recognition dataset that has been widely used as benchmark for SET[2]. It includes 2000 5s-long isolated sound recordings from 50 sound classes. The dataset covers a large variety of sound classes that range from domestic and urban sounds, over nature and animal sounds, to human non-speech sounds. As all DAEs are pre-trained on the AudioSet dataset, we consider the audio clips of the ESC-50 dataset as previously unseen and hence as suitable data for our experiments.

Our research focus is on embedding space analysis for SET. To this end, we create 1000 random pairs of sounds taken from the ESC-50 dataset. Given our random sound assignment, the large majority of 98.2% of the sound pairs include sounds from different classes. From each sound pair, we create six mixtures by blending between the sound pairs. In addition to its unprocessed version, we create four degraded versions of each mixture using the methods listed in in Table II. These versions are used in Section VI to study the robustness of different audio representations towards audio degredations. We refer to this dataset as "ESC50Mix" in the following.[3]

For the m-th random sound pair $(x_{m,1}, x_{m,2})$ with $x_{m,1} \in \mathbb{R}^{L \cdot f_\mathrm{s}}$, $x_{m,2} \in \mathbb{R}^{L \cdot f_\mathrm{s}}$ with $L = 5$ and $m \in [1:1000]$, we create sound mixtures using a mixture coefficient $\gamma_g \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ (indexed by $g \in \{1:6\}$) and apply a degradation function Λ_a (indexed by $a \in \{1:5\}$) as

$$x_{m,g,a} = \Lambda_a \left(\gamma_g \cdot x_{m,1} + (1 - \gamma_g) \cdot x_{m,2} \right) \tag{1}$$

with m denoting the sound pair index. The mixing indices $g \in \{1, 6\}$ result in the isolated sounds $x_{m,2}$ and $x_{m,1}$, respectively, while $g \in [2:5]$ result in mixtures of both sounds. No normalization is applied to the original ESC-50 audio clips.

[2][Online]. Available: https://github.com/karolpiczak/ESC-50

[3]The dataset has been published at https://zenodo.org/record/7913031

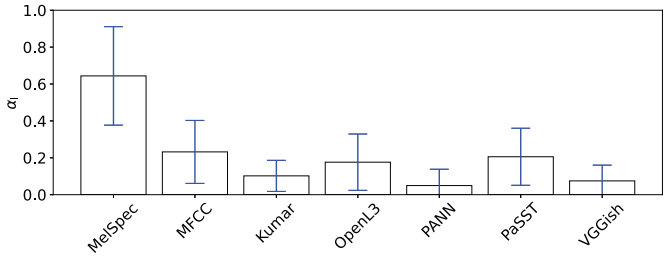

Fig. 3. Inter-correlation strength α_{i} for different audio representations. Error bars indicate standard deviation across feature dimension pairs.

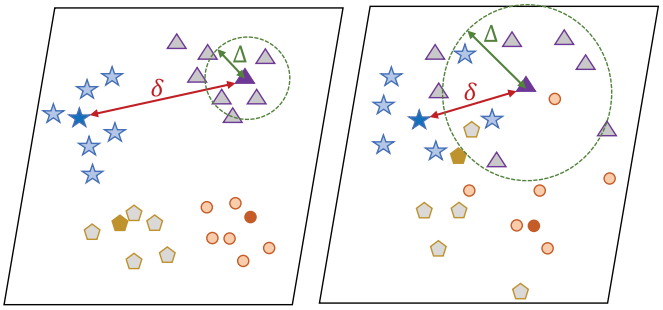

Fig. 4. Possible embedding space configurations with four well-separated classes (left) and four partially overlapping classes (right). The intracluster distance Δ for the purple class as well as the intercluster distance δ between the purple and blue classes are shown as examples.

As detailed in Table II, we use the `audiomentations`[4] Python library to implement loudness reduction, additive Gaussian noise, as well as low-frequency and high-frequency boost as degradation functions Λ_a. In total, the ESC50Mix dataset includes 30000 audio clips. Given an embedding function $f(\cdot)$, each sound mixture $x_{m,g,a}$ is mapped to its corresponding embedding vector $z_{m,g,a} \in \mathbb{R}^E$ as

$$z_{m,g,a} = f(x_{m,g,a}). \tag{2}$$

For the experiments conducted in this article, we stack the embedding vectors of all 1000 sound pairs separately for each combination of embedding function and audio degradation method as an embedding matrix $Z \in \mathbb{R}^{1000\times E}$.

V. Embedding Space Exploration

In this section, we introduce several measures to explore the embedding space distributions of the audio representations introduced in Section III.

A. Inter-Correlation

We investigate the redundancy of an audio representation by measuring the average pair-wise correlation between its feature dimensions. We use the sample Pearson correlation coefficient as correlation measure. It is defined as

$$r(q,v) = \frac{\sum_{i=1}^{K}(q_i-\mu_q)(v_i-\mu_v)}{\sqrt{\sum_{i=1}^{K}(q_i-\mu_q)^2}\sqrt{\sum_{i=1}^{K}(v_i-\mu_v)^2}} \tag{3}$$

for two vectors $q \in \mathbb{R}^K$ and $v \in \mathbb{R}^K$ with $K \in \mathbb{Z}$ and their means $\mu_q \in \mathbb{R}$ and $\mu_v \in \mathbb{R}$, respectively.

Given a stacked embedding matrix $Z \in \mathbb{R}^{N\times E}$ of N row-wise stacked embedding vectors $z \in \mathbb{R}^E$, we measure the inter-correlation strength α_{i} as

$$\alpha_{\mathrm{i}} = \frac{1}{E(E-1)} \sum_{i\in[1:E]} \sum_{\substack{j\in[1:E]\\ j=i}} |r(Z[:,i], Z[:,j])| \tag{4}$$

with $Z[:,i] \in \mathbb{R}^N$ denoting the i-th column of Z. For simplicity, we only investigate the non-degraded isolated sound recordings ($a = 1, g \in \{1,6\}$).

As shown in Fig. 3, DAEs are less redundant audio representations than NTARs as their inter-correlation strength is lower. This holds in particular for the DAEs trained in a supervised fashion (`Kumar`, `PANN`, and `VGGish`). The high value for `MelSpec` is expectable since adjacent filters in the triangular filterbank for the Mel-frequency mapping overlap. `MFCC`, `OpenL3`, and `PaSST` show similar inter-correlation strength values.

[4]https://github.com/iver56/audiomentations

B. Class Scattering

If we consider a multi-class classification scenario, an ideal embedding space has dense and well-separated clusters for each class as shown on the left side of Fig. 4. In contrast, as shown on the right side, partially overlapping classes can cause confusions between adjacent classes and hence complicate the task for a subsequent classification model. In this section, we use the Dunn Index (DI) [49] and the Davies-Bouldin Index (DBI) [50] as two established separation measures to characterize the class scattering in the embedding space for the non-degraded isolated sound recordings ($a = 1, g \in \{1,6\}$).

Given a set $\mathcal{Z} = \{z \in \mathbb{R}^E\}$ of embedding vectors of isolated sounds, let $\mathcal{Z}_i = \{z \in \mathcal{Z} \mid c(z) = i\}$ be the subset of all embedding vectors labeled with class $c(z) \in [1:C]$ where $C \in \mathbb{N}$ denotes the number of classes. The class centroids in the embedding space are computed as

$$\mu_i = \frac{1}{|\mathcal{Z}_i|} \sum_{z\in\mathcal{Z}_i} z. \tag{5}$$

The intracluster distance Δ_i measures the average distance of all class samples to their class centroid as

$$\Delta_i = \frac{1}{|\mathcal{Z}_i|} \sum_{z\in\mathcal{Z}_i} d(z,\mu_i) \tag{6}$$

where $d(\cdot)$ denotes the Euclidean distance. The intercluster distance $\delta_{i,j}$ measures the distance between the two class centroids of class i and j as

$$\delta_{i,j} = d(\mu_i,\mu_j). \tag{7}$$

Based on these concepts, the Dunn Index α_{DI} looks for the closest pair of clusters as well as the most spread cluster to derive a

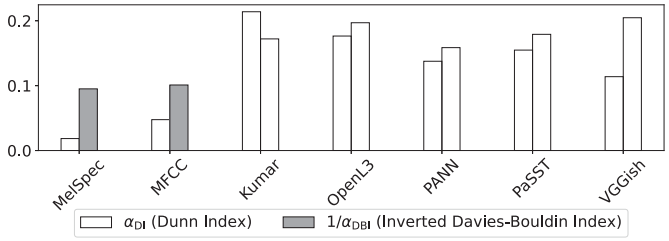

Fig. 5. Dunn Index (DI) and Inverted Davies-Bouldin Index (DBI) for different audio representations. Higher values indicate a better class separability.

separation measure as

$$\alpha_{\text{DI}} = \frac{\min_{\substack{i,j\in[1:C]\\ i=j}} \delta_{i,j}}{\max_{i\in[1:C]} \Delta_i} \tag{8}$$

This measure follows a "pessimistic" view and only considers the most poorly segregated and widely dispersed classes.

As a second measure, the Davies-Bouldin Index first computes pair-wise cluster similarity measures as

$$R_{i,j} = \frac{\Delta_i + \Delta_j}{\delta_{i,j}} \tag{9}$$

Then, for each class, the most similar other class is identified and these similarity values are finally averaged as

$$\alpha_{\text{DBI}} = \frac{1}{C} \sum_{i\in[1:C]} \max_{\substack{j\in[1:C]\\ j=i}} R_{i,j} \tag{10}$$

Since α_{DBI} decreases with improved class separability, we use the inverted DBI measure ($1/\alpha_{\text{DBI}}$) as a separation measure of a given clustering.

Fig. 5 summarizes the two measures for all audio representations. Both measures show a similar trend that DAEs in general yield a better class scattering in the embedding space. At the same time, there is no clear evidence whether DAEs trained in a supervised or self-supervised fashion show a better separability.

VI. Sensitivity to Audio Degradations

In this section, we aim to measure the sensitivity of audio representations against different types of audio degradations. Such degradations are caused by acoustic variations such as background noise and loudness variations, which often appear in real-world sound monitoring scenarios. Ideally, an embedding function $f(\cdot)$ is robust against such audio degradations since they do not change the semantics of the sound classes to be recognized.

As illustrated in Fig. 6, we consider two types of sensitivity measures. In the class-agnostic measure ψ_{a} (see left side of Fig. 6), we do not take the class membership of embedding vectors into account. Instead, we consider only the distance between the non-degraded embedding vector z and the degraded embedding vector z_{d}:

$$\psi_{\text{a}} = \frac{1}{|\mathcal{Z}|} \sum_{z\in\mathcal{Z}} d(z, z_{\text{d}}). \tag{11}$$

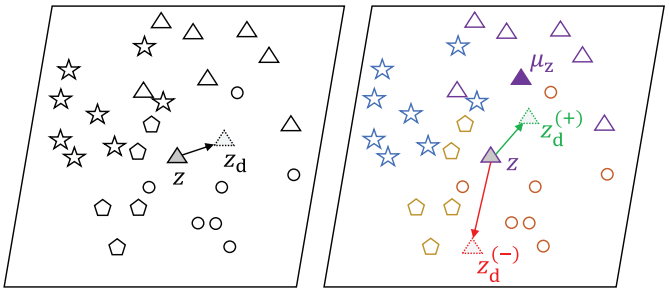

Fig. 6. Class-agnostic view (left side) and class-centric view (right side) for computing the sensitivity towards a degradation function, which causes a embedding vector z to move to z_{d}. For the class-centric view, $z_{\text{d}}^{(+)}$ and $z_{\text{d}}^{(-)}$ show two cases where the embedding vector moves either towards or away from its corresponding class centroid μ_z.

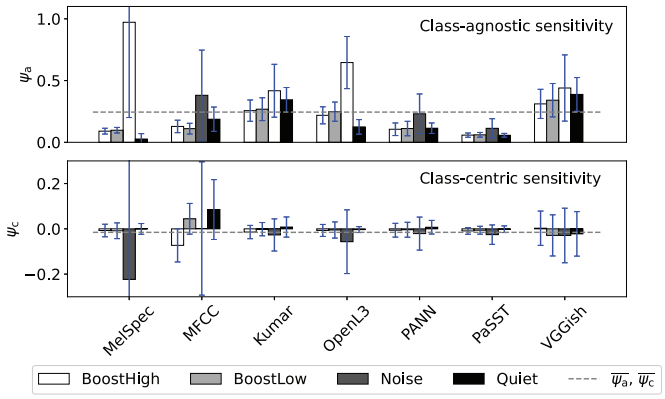

Fig. 7. Sensitivity measures ψ_{a} (upper plot) and ψ_{c} (lower plot) observed per audio representation and degradation type. Error bars indicate standard deviation across test samples. Horizontal dashed lines indicate global averages $\overline{\psi_{\text{a}}}$ and $\overline{\psi_{\text{c}}}$ over both sensitivity values.

In the class-centric measure ψ_{c} (see right side of Fig. 6), we measure the "drift" of an embedding vector relative to its corresponding class centroid $\mu_{c(z)}$:

$$\psi_{\text{c}} = \frac{1}{|\mathcal{Z}|} \sum_{z\in\mathcal{Z}} \big(d(z, \mu_{c(z)}) - d(z_{\text{d}}, \mu_{c(z)})\big). \tag{12}$$

Positive values for ψ_{c} indicate that an embedding vector moves towards its class centroid whereas negative values indicate a drift away from it.

In our experiment, we analyze the degraded versions of the isolated sound recordings in the ESC50Mix dataset ($g \in \{1, 6\}$). Fig. 7 illustrates the mean sensitivity values and the corresponding error bars based on the standard deviation computed over all audio recordings. The class-agnostic sensitivity measure ψ_{a} show that the "Noise" degradation has the strongest impact on the embeddings and generally causes the embeddings to move away from their class centroids with the exception of `MFCC`, where ψ_{c} remains almost zero on average. This is expectable, as the MFCC provide a decorrelated approximation of the spectral envelope, which naturally suppresses noise.

On the other hand, the "Quiet" degredation, which reduces the loudness, leads to an embedding drift for all representations except for `MelSpec`. Notably, the `MFCC`s seem robust to such

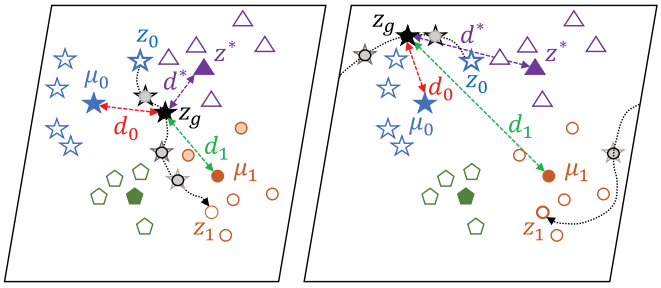

Fig. 8. Two possible embedding space trajectories that are obtained by blending between two sounds z_0 and z_1. Given a sound mixture z_g of these sounds, the plots show the distances d_0, d_1, and d^* towards the corresponding class centroids μ_0 and μ_1 as well as the closest out-of-class centroid μ^*. The trajectory (black dotted line) indicates a "wrapped" continuous path on a submanifold embedded in a high-dimensional feature space.

degradation as they tend to drift towards their class centroids which does not introduce a higher risk for misclassifications. The two audio degradations "BoostHigh" and "BoostLow", which alter the spectral envelope, have a stronger effect on the DAEs than on the NTARs. However, both do not cause a strong drift towards or away from the class centroids. In summary, while this investigation revealed individual strengths and weaknesses of all embeddings, the PaSST as well as the PANN embeddings appear to be in overall the most robust representations against the investigated audio degradations.

VII. Sound Blending Trajectories

In this experiment, we investigate how a blending between two isolated sounds maps to a trajectory between their embedding vectors. We argue that if such a trajectory passes by other classes in the embedding space, misclassification can be caused. Fig. 8 illustrates this idea: Given the embedding vectors z_0 and z_1 of two isolated sounds, we investigate the trajectory corresponding to the embedding vectors of the mixtures z_g of both sounds, which according to (1) depends on the mixing coefficient g. In this experiment, we do not apply any audio degradation ($a = 1$).

A. Embedding Space Distances

Given an example mixture z_g along this trajectory, we measure its embedding space distance to the class centroids μ_0 and μ_1 of both isolated sounds as $d_0 = d(z_g, \mu_0)$ and $d_1 = d(z_g, \mu_1)$ as well as its distance to the closest out-of-class centroid μ_z^* as $d^* = d(z_g, \mu_z^*)$. Fig. 8 illustrates two possible trajectories: The first trajectory (left plot) passes by two other classes (purple triangles, green pentagons) and shows potential for sound misclassification. The second trajectory (right plot) remains close to the original classes (blue stars, orange circles) and the mixtures remain further away from out-of-class centroids.

Fig. 9 shows the dependence of the three distance values d_0, d_1 and d^* on the mixing parameter g for different audio representations. The plots show the mean values averaged over all 1000 sound pairs to illustrate general trends. We make several observations: First, d^* has a convex shape and is consistently below d_1 and d_2 for the NTARs as well as for OpenL3. This

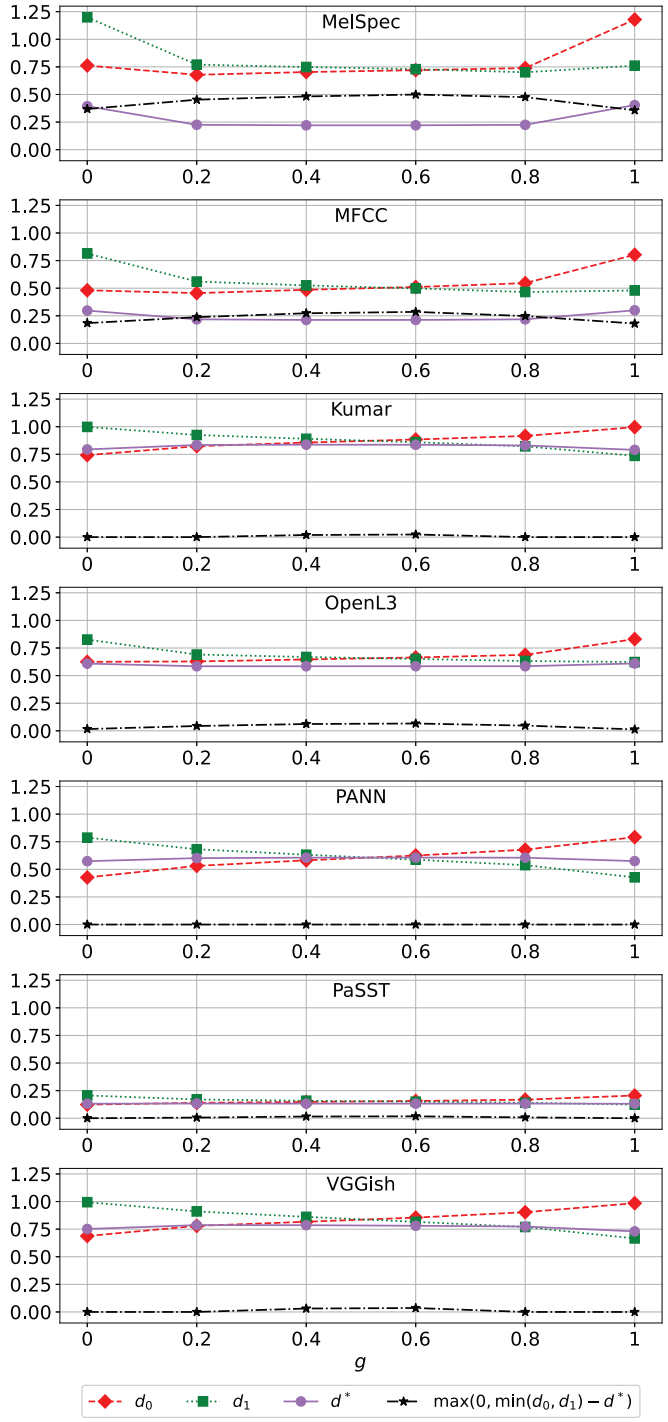

Fig. 9. Subplots show for all embedding types the distances d_0 and d_1 between the mixture embeddings z to the class centroids $\mu_{z,0}$ and $\mu_{z,1}$ of the corresponding sound classes as well as the distance d^* to the closest out-of-class centroid μ_z^*. A likelihood measure for class confusion is derived as $\min(d_0, d_1) - d^*$ and shown here only for positive values.

property is disadvantageous for SET as it indicates that both the isolated sounds as well as the mixtures tend to remain closer to out-of-class centroids than to their corresponding class centroids. When looking at the supervised DAEs (Kumar, PANN, PaSST, and VGGish), d^* has a concave shape, which indicates that for stronger mixtures ($g \in \{0.4, 0.6\}$), the

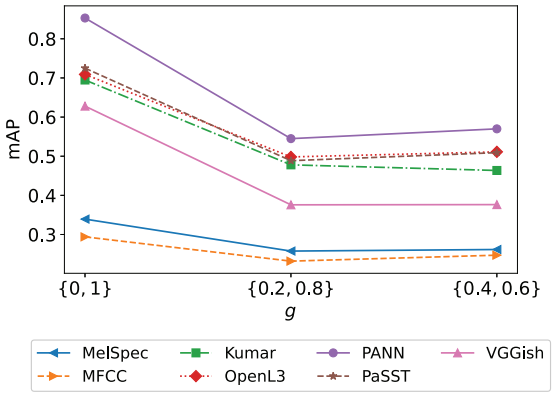

Fig. 10. Multi-label SET performance of a two-layer MLP model measured as macro-level mean average precision (mAP), which is shown for different embeddings and different types of sound mixtures based on the mixing coefficient g.

potential for confusion with other classes is smaller than for isolated sounds ($g \in \{0, 1\}$) or soft mixtures ($g \in \{0.2, 0.8\}$). We further illustrate in Fig. 9 a likelihood measure for class confusion derived as $\max(0, \min(d_0, d_1) - d^*)$. In particular for the PANN and PaSST embeddings, we have the desired property that $\min(d_0, d_1) < d^*$ holds true for all mixture coefficients g. Therefore, we expect the structure of the embedding spaces of both the PANN and PaSST embeddings to be most suitable for low-polyphony SET among the investigated audio representations as class confusions due to proximate out-of-class centroids being mostly avoided.

B. SET Experiment

Complementary to the embedding space distance investigations presented in Section VII-A, we run a SET experiment using the ESC50Mix dataset. We use the first 800 sound pairs as training data and the last 200 sound pairs as test data. Again, we focus on the non-degraded audio clips ($a = 1$) and consider all gain factors g when compiling the training and test datasets. Consequently, all SET models are trained and evaluated with both single-label and multi-label audio clips. In particular, we obtain single-label annotations for all audio clips with only one sound being audible ($\gamma_g = 0$ and $\gamma_g = 1$) and multi-label annotations for all sound mixtures.

Inspired by [7], we use a two-layer Multi-Layer Perceptron (MLP) model as a classifier to process the embeddings, which consists of a first layer with 128 neurons and a Rectified Linear Unit (ReLU) activation function and a second layer of 50 neurons with a sigmoid activation function. All models are trained for 150 epochs using binary crossentropy loss, the Adam optimizer [51] with a learning rate of 10^{-3}, and a batch size of 32. We randomly use 20 % of the training set as validation set and use early stopping on the validation loss to stop the training. The macro-average mean average precision (mAP) is computed as evaluation metric.

Fig. 10 summarizes the mAP values obtained for three types of sound mixtures ranging from isolated sounds ($g \in \{0, 1\}$) over mixtures of one predominant and one background sound ($g \in \{0.2, 0.8\}$) to mixtures of two sounds of similar intensity ($g \in \{0.4, 0.6\}$). It comes by no surprise that we can observe a decrease in mAP from single-label audio clips with isolated sounds to multi-label audio clips. Interestingly, the tagging models perform slightly better for the multi-label clips when both sounds have a similar intensity. The NTARs perform significantly worse than the DAEs. Presumably, the shallow MLP model is less expressive using NTARs, which characterize sounds only by the shape of their spectral envelopes. DAEs, in contrast, are trained to capture more complex temporal-spectral patterns. When comparing the different DAEs, the PANN embedding perform best followed by OpenL3 and PaSST embeddings, which perform en par. This confirms our findings from Section VII-A, where DAEs clearly outperformed the NTARs. As only exception, the OpenL3 embeddings perform better than expected based on the observed embedding space distance relationships.

VIII. Conclusion

Motivated by the challenges of deploying machine listening approaches for real-life application scenarios, we study in this article the suitability of two non-trainable audio representations (NTARs) as well as five deep audio embeddings (DAEs) for SET. We first investigated general properties of these embeddings such as the redundancy caused by feature inter-correlations as well as the class separability in the embedding spaces. Then, we assessed the robustness of the embeddings against four types of audio degredations. We proposed two measures based on a class-agnostic and a class-centric view on the resulting embedding drift in the embedding space. We observed that both NTARs and DAEs have individual weaknesses while the PaSST and PANN embeddings seem to be the most robust representations.

As a main contribution of this article, we blended between random sound pairs to create sound mixtures and studied the resulting embedding space trajectories to assess the risk of sound misclassification. This arises if sound mixtures are too close to other sound classes in the embedding space. Again, we found that the embedding space of the PANN embeddings seems to be structured in such way that sound mixtures generally remain close enough to their original sound classes, which leads to superior SET performance. Our analyses so far are based on a low sound polyphony of two overlapping sounds. As future work, new training approaches should be developed to learn DAEs which better account for sound mixtures of higher polyphony degrees, which are common in real-world soundscapes. Another open question is, how the proposed embedding space trajectories can be generalized to higher sound polyphony degrees with a rapidly increasing number of sound permutations.

Acknowledgment

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer Institute for Integrated Circuits IIS.

REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. E. Ellis, *Computational Analysis of Sound Scenes and Events*, 1st ed. Berlin, Germany: Springer, 2018.

[2] S. Hershey et al., "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 131–135.

[3] A. Kumar, M. Khadkevich, and C. Fügen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 326–330.

[4] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 3852–3856.

[5] 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, 2020.

[6] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. Conf. Int. Speech Commun. Assoc.*, Incheon, Korea, 2022, pp. 2753–2757.

[7] S. Grollmisch, E. Cano, C. Kehling, and M. Taenzer, "Analyzing the potential of pre-trained embeddings for audio classification tasks," in *Proc. IEEE 28th Eur. Signal Process. Conf.*, 2020, pp. 790–794.

[8] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, and M. D. Plumbley, "Source separation with weakly labelled data: An approach to computational auditory scene analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 101–105.

[9] Y. Hou, B. Kang, W. V. Hauwermeiren, and D. Botteldooren, "Relation-guided acoustic scene classification aided with event embeddings," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2022, pp. 1–8.

[10] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE Proc. IRE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[11] S. Liu et al., "Audio self-supervised learning: A survey, 2022, *arXiv:2203.01205*.

[12] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Toronto, ON, Canada, 2021, pp. 3875–3879, doi: 10.1109/ICASSP39728.2021.9413528.

[13] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 4563–4567.

[14] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[15] S. Nawaz, M. K. Janjua, I. Gallo, A. Mahmood, and A. Calefati, "Deep latent space learning for cross-modal mapping of audio and visual signals," in *Proc. IEEE Int. Conf. Digit. Image Comput.: Techn. Appl.*, Perth, Australia, 2019, pp. 1–7.

[16] Q. Wang et al., "A model ensemble approach for audio-visual scene classification," Detection and Classification of Acoustic Scenes and Events, Tech. Rep., pp. 1–5, 2021. [Online]. Available: https://dcase.community/documents/challenge2021/technical_reports/DCASE2021_Du_124_t1.pdf2021.

[17] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 640–649.

[18] M. Won, J. Salamon, N. J. Bryan, G. J. Mysore, and X. Serra, "Emotion embedding spaces for matching music to stories," in *Proc. Soc. Music Inf. Retrieval Conf.*, Online, 2021, pp. 777–785.

[19] Y.-N. Hung and A. Lerch, "Feature-informed embedding space regularization for audio classification," in *Proc. 30th Eur. Signal Process. Conf.*, 2022, pp. 419–423.

[20] P. Lopez-Meyer, J. A. del Hoyo Ontiveros, H. Lu, and G. Stemmer, "Efficient end-to-end audio embeddings generation for audio classification on target applications," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 601–605.

[21] J. Engel et al., "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proc. Int. Conf. Mach. Learn.*, Sydney, Australia, 2017, pp. 1068–1077.

[22] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A. generative model for music, 2020, *arXiv:2005.00341*.

[23] A. Natsiou and S. O'Leary, "Audio representations for deep learning in sound synthesis: A review," in *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl.*, Tangier, Morocco, 2021, pp. 1–8.

[24] P. Esling, A. Chemla–Romeu-Santos, and A. Bitto, "Generative timbre spaces: Regularizing variational auto-encoders with perceptual metrics," in *Proc. Int. Conf. Digit. Audio Effects*, Aveiro, Portugal, 2018, pp. 1–8.

[25] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, New Orleans, LA, USA, 2017, pp. 776–780.

[26] T. Iqbal, Y. Cao, M. D. Plumbley, and W. Wang, "Incorporating auxiliary data for urban sound tagging," Detection and Classification of Acoustic Scenes and Events, Tech. Rep., pp. 1–4, 2020. [Online]. Available: https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Iqbal_38_t5.pdf

[27] S. Shishkin, D. Hollosi, S. Dolco, and S. Goetze, "Active learning for sound event classification using monte-carlo dropout and PANN embeddings," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, Online, 2021, pp. 150–154.

[28] Y. Wang, N. J. Bryan, J. Salamon, M. Cartwright, and J. P. Bello, "Who calls the shots? Rethinking few-shot learning for audio," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, New Paltz, NY, USA, 2021, pp. 36–40.

[29] A. O. Eren and M. Sert, "Audio captioning based on combined audio and semantic embeddings," in *Proc. IEEE Int. Symp. Multimedia*, Naples, Italy, 2020, pp. 41–48.

[30] Y. Zhang et al., "ACTUAL: Audio captioning with caption feature space regularization", *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, to be published, doi: 10.1109/TASLP.2023.3293015.

[31] S. Braun and H. Gamper, "Effect of noise suppression losses on speech distortion and ASR performance," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 996–1000.

[32] D. Ngo, L. Pham, T. Hoang, S. Kolozali, and D. Jarchi, "Audio-based deep learning frameworks for detecting COVID-19," in *Proc. IEEE 30th Eur. Signal Process. Conf.*, 2022, pp. 1233–1237.

[33] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Zero-shot audio source separation through query-based learning from weakly-labeled data," in *Proc. Conf. Artif. Intell.*, Virtual, 2022, pp. 4441–4449.

[34] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 7118–7122.

[35] J. Turian et al., "HEAR: Holistic evaluation of audio representations," in *Proc. NeurIPS Competitions Demonstrations Track*, 2022, pp. 125–145.

[36] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 1, pp. 208–221, Jan. 2017.

[37] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, "Measuring domain shift for deep learning in histopathology," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 2, pp. 325–336, Feb. 2021.

[38] S. Srivastava et al., "A study on robustness to perturbations for representations of environmental sound," in *Proc. IEEE Eur. Signal Process. Conf.*, Belgrade, Serbia, 2022, pp. 125–129.

[39] I. Kukanov, V. Hautamäki, and K. A. Lee, "Maximal figure-of-merit embedding for multi-label audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 136–140.

[40] Z. Li, M. Mozer, and J. Whitehill, "Compositional embeddings for multi-label one-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 296–304.

[41] M. Casey, C. Rhodes, and M. Slaney, "Analysis of minimum distances in high-dimensional musical spaces," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 1015–1028, Jul. 2008.

[42] B. McFeeet al., "Librosa/Librosa: 0.8.0," 2020. Accessed: Aug. 2022 [Online]. Available: https://zenodo.org/record/3955228

[43] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50 K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 30, pp. 829–852, 2022.

[44] R. Arandjelović and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 609–617.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015, pp. 1–14.

[46] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 571–575.

[47] D. S. Park et al., "SpecAugment: A simple augmentation method for automatic speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 2613–2617.

[48] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23 rd Annu. ACM Conf. Multimedia*, Brisbane, Australia, 2015, pp. 1015–1018.
[49] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973.
[50] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI- 1, no. 2, pp. 224–227, Apr. 1979.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, San Diego, USA, 2015. [Online]. Available: https://dare.uva.nl/search?identifier=a20791d3-1aff-464a-8544-268383c33a75

Jakob Abeßer (Member, IEEE) received the Diploma in computer engineering from the Technische Universität Ilmenau, Ilmenau, Germany, in 2008. In 2014, he received the Ph.D. degree in media technology. Since 2008, he has been a Research Scientist with the Fraunhofer Institute for Digital Media Technology (IDMT), Ilmenau. In 2010, he was with the Centre of Excellence in music, mind, body and brain, University of Jyväskylä, Jyväskylä, Finland, for a research stay. His dissertation, supervised by Prof. Gerald Schuller, deals with automatic transcription, classification, and synthesis of bass guitar recordings. From 2012 to 2017, he further joined the University of Music Franz Liszt, 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, a Senior Scientist with Fraunhofer IDMT. He is also currently a Visiting Researcher with the Semantic Audio Processing Group headed by Prof. Meinard Müller with International Audio Laboratories, Erlangen, Germany. Working towards a habilitation degree, his research interests include intersection between machine listening and music information retrieval.

Sascha Grollmisch received the engineering diploma in Media Technology in 2009 from Technische Universität Ilmenau, Illmenau. He started his career as a Software Developer with Fraunhofer Institute for Digital Media Technology (IDMT). Motivated by his passion for music, he was later part of the spin-off company Songquito, which distributes the music education software Songs2See, developed within a long-term research project with Fraunhofer IDMT. For their effort in developing one of the first fully interactive music learning games, the Songquito team was the recipient of 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 working toward the Ph.D. degree with 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.

Meinard Müller (Fellow, IEEE) received the Diploma in mathematics and Ph.D. degree in computer science from the University of Bonn, Bonn, Germany, in 1997 and 2001, respectively. After the Postdoctoral studies (2001-2003) in Japan and Habilitation (2003-2007) in multimedia retrieval in Bonn, he was a Senior Researcher with Saarland University, Saarbrücken, Germany, and the Max-Planck Institut für Informatik (2007-2012), Saarbrücken, Germany. Since 2012, he has held a Professorship for Semantic Audio Signal Processing with the International Audio Laboratories Erlangen, Erlangen, Germany, a joint institute of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer Institute for Integrated Circuits IIS. His research interests include music processing, music information retrieval, audio signal processing, and motion processing. He was a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee (2010-2015), Member of the Senior Editorial Board of the IEEE Signal Processing Magazine (2018-2022), and Member of the Board of Directors, International Society for Music Information Retrieval (2009-2021, being its President in 2020/2021). In 2020, he was elevated to IEEE Fellow for contributions to music signal processing.