

HEAR 2021: Holistic Evaluation of Audio Representations

Joseph Turian	TURIAN@GMAIL.COM
Jordie Shier	JSHIER@UVIC.CA
Humair Raj Khan	KHUMAIRRAJ@GMAIL.COM
Bhiksha Raj	BHIKSHA@CS.CMU.EDU
Björn W. Schuller	BJOERN.SCHULLER@IMPERIAL.AC.UK
Christian J. Steinmetz	C.J.STEINMETZ@QMUL.AC.UK
Colin Malloy	MALLOYC@UVIC.CA
George Tzanetakis	GTZAN@IEEE.ORG
Gissel Velarde	GV@URUBO.ORG
Kirk McNally	KMCNALLY@UVIC.CA
Max Henry	MAX.HENRY@MAIL.MCGILL.CA
Nicolas Pinto	NICOLAS.PINTO@GMAIL.COM
Camille Noufi	CNOUFI@STANFORD.EDU
Christian Clough	CHRISTIAN.CLOUGH@GMAIL.COM
Dorien Herremans	DORIEN.HERREMANS@GMAIL.COM
Eduardo Fonseca	EDUARDO.FONSECA@UPF.EDU
Jesse Engel	JESSEENGEL@GOOGLE.COM
Justin Salamon	SALAMON@ADOBE.COM
Philippe Esling	PHILIPPE.ESLING@IRCAM.FR
Pranay Manocha	PMANOCHA@PRINCETON.EDU
Shinji Watanabe	SWATANAB@ANDREW.CMU.EDU
Zeyu Jin	ZEJIN@ADOBE.COM
Yonatan Bisk	YBISK@CS.CMU.EDU

Editor: Editor’s name

Abstract

What audio embedding approach generalizes best to a wide range of downstream tasks across a variety of everyday domains without fine-tuning? The aim of the HEAR 2021 NeurIPS challenge is to develop a general-purpose audio representation that provides a strong basis for learning in a wide variety of tasks and scenarios. HEAR 2021 evaluates audio representations using a benchmark suite across a variety of domains, including speech, environmental sound, and music. In the spirit of shared exchange, each participant submitted an audio embedding model following a common API that is general-purpose, open-source, and freely available to use. Twenty-nine models by thirteen external teams were evaluated on nineteen diverse downstream tasks derived from sixteen datasets. Open evaluation code, submitted models and datasets are key contributions, enabling comprehensive and reproducible evaluation, as well as previously impossible longitudinal studies. It still remains an open question whether one single general-purpose audio representation can perform as holistically as the human ear.

Keywords: audio representations, representation learning, embeddings, transfer learning, multi-task learning, multi-modal learning, classification, tagging

1. Introduction

The codification of strong general-purpose representations in natural language and computer vision has led to a renaissance in multimodal modeling and increased cross-discipline collaboration. Audio is an equally rich source of information about the world, but outside of speech recognition it has not achieved the same degree of attention from the machine learning community. This is a key challenge for the community, as good representations support good machine learning. And robust evaluation enables general representations. Broad evaluation suites help prevent overfitting to common test sets (Recht et al., 2018) and have improved the state-of-the-art on language and vision representation learning (Wang et al., 2019b,a; Goyal et al., 2019; Zhai et al., 2019; DeYoung et al., 2020). In general practice, audio representations are not evaluated on a broad range of audio problems, and as a result, it is difficult to know which audio representation to use for a novel audio learning task.

The Holistic Evaluation of Audio Representations (HEAR) challenge was created to encourage the development of flexible audio representations, to give greater insight into how audio representations will generalize, and to enable fast development cycles both for researchers developing new models and researchers applying existing models. HEAR participants submitted audio representation models that are general-purpose, open-source, and freely available to use off-the-shelf. All HEAR compatible models follow a common API, which makes switching between models as simple as changing one line of code.

The HEAR benchmark includes nineteen tasks: five open tasks derived from three datasets for which the problem definition and evaluation data were available to participants, and 14 additional secret tasks for evaluation, to which participants were completely blind throughout the challenge. While most of the tasks (open or secret) have good or promising solutions when worked on in isolation, the novelty of this challenge is that the *same* representation must be used to solve all of them. These tasks encompass multiple audio domains: speech, environmental sound, and music, with tasks that involve short and long time spans. HEAR datasets are easy to use: all are preprocessed to a common format with standard splits and self-explanatory human-readable metadata, and are distributed as tarfiles online.¹ This alleviates the engineering effort required to work with datasets that require YouTube scraping, have variably documented preprocessing requirements, or are gatekept through closed-access request forms. Researchers are also welcome to use HEAR datasets under entirely open licenses (many of which allow commercial use), without using our downstream evaluation code.

Evaluation consists of classification tasks, both multiclass and multilabel, requiring either prediction over the entire audio scene (clip), or temporal-based onset detection of sound event (Mesaros et al., 2016). HEAR-compatible models can generate an embedding of arbitrary size, which is fed into a simple generic predictor by our open-source evaluation algorithm. Evaluation code, submitted models, and datasets are all available at <https://neuralaudio.ai/hear.html>.

1. <https://zenodo.org/record/5885750>

2. Background on representation learning

At a high level, a learned representation (embedding) consists of a machine learning model that takes a low-level representation of the input and outputs a numerical representation, typically a fixed-size vector, that lends itself well to discriminative tasks (e.g., by training a simple MLP on these embeddings). A good representation should (1) transfer to a wide range of different tasks and (2) transfer with limited supervision (Goyal et al., 2019, 2022). The scope of HEAR is limited to discriminative tasks.

In the following paragraphs, we describe trends from the natural language processing (NLP) and vision literature on representation learning, some of which have been applied to audio. Vision work is particularly relevant (Amiriparian et al., 2017), as 2-D transformations of audio, such as the widely used log-Mel spectrogram (Davis and Mermelstein, 1980), lend themselves well to methods designed to process 2-D input data. For this reason, a common thread in the literature on audio representations is that vision models are applied to 2-D audio representations. With that said, many of the insights from text-based language modeling, such as autoregressive neural modeling (Bengio et al., 2001), predicting tokens masking as an unsupervised pretext task (Collobert et al., 2011), and bidirectional transformers (Devlin et al., 2019), have found their way into the audio literature, e.g., WaveNet (van den Oord et al., 2016), wav2vec (Schneider et al., 2019), and HuBERT (Hsu et al., 2021), respectively. Textless NLP like Generative Spoken Language Modeling (GSLM, Lakhotia et al. (2021)), applies an NLP lens to spoken audio instead of written text.

Inducing representations The shallowest representation for audio is the raw digital audio signal itself. However, its extremely high dimensionality means it is rarely useful for discriminative tasks without additional processing, whether via manually crafted DSP engineering or transformations learned by training a neural network (Trigeorgis et al., 2016). Better representations might be obtained by applying a hand-crafted transformation based upon domain-expertise, such as the log-scaled Mel spectrogram (Davis and Mermelstein, 1980), Mel Frequency Cepstral Coefficients (MFCC, Logan (2000)), constant Q-transform (Schörkhuber and Klapuri, 2010), or scattering transform (Andén and Mallat, 2014). Audio filterbanks can also be learned (Zeghidour et al., 2021). Deep ML architectures can extract even more abstract, high-level representations (Aytar et al., 2016; Hershey et al., 2017; Cramer et al., 2019). Purely randomly weighted architectures impose particular inductive biases on data and can do better than hand-crafted baselines (Saxe et al., 2011; Pons and Serra, 2019). However, it is more common to train these architectures.

Architectures The architecture of the model typically includes an encoder to transform the input, and can optionally also include temporal modelling to capture context, and/or a generative decoder. A common encoder architecture uses Convolutional Neural Networks (CNN) applied to a 2-D input (Hershey et al., 2017; Cramer et al., 2019), or directly to the 1-D audio signal (van den Oord et al., 2016; Baevski et al., 2020). Temporal context modelling is often achieved via Recurrent Neural Networks (RNN) (Merhi et al., 2017; Kalchbrenner et al., 2018), or Transformers (Baevski et al., 2020). The latter, in particular, have achieved strong results for audio classification (Gong et al., 2021a), though they are costly to train from scratch. Koutini et al. (2021) (§4) demonstrate a faster training approach for audio transformer, which requires two GPU-days to pretrain on AudioSet. In reaction to the use

of transformers, all-MLP architectures have demonstrated competitive results on language and vision tasks (Liu et al., 2021; Tolstikhin et al., 2021).

Training regimes Models can be trained on a (large-scale) supervised task, such as ImageNet (Deng et al., 2009) for vision and AudioSet (Gemmeke et al., 2017) for audio. Multitask supervised training can further improve generalization (Pascual et al., 2019).

To avoid the need for human-labeling, self-supervised models (a form of unsupervised learning) learn from large-scale unlabeled corpora. Many self-supervised approaches learn to correspond the original input with a different view on that same input, such as a semantically identical augmentation (Chen et al., 2020b; Tian et al., 2020). To avoid collapsed solutions, self-supervised approaches historically used negative samples with a triplet loss (Chopra et al., 2005), possibly requiring large negative batches (Chen et al., 2020b; Saeed et al., 2021), which can be expensive to train. Alternatives include quantization approaches to define uniform clusterings of representations (Baevski et al., 2020), or carefully implemented asymmetric training architectures like BYOL (Grill et al., 2020; Niizumi et al., 2021) and SimSiam (Chen and He, 2021). More recent are self-supervised approaches that avoid these aforementioned techniques, relying instead upon explicit and fundamental priors (Zbontar et al., 2021; Bardes et al., 2022). Input augmentations can be used to increase the size of training data or provide corresponding views on the input (Salamon and Bello, 2017). Fonseca et al. (2021b) and Wang and van den Oord (2021) discuss augmentations, including audio mixing, which Gong et al. (2021b); Wang et al. (2021b) explore in greater depth and argue is useful both for supervised and unsupervised regimes.

Multi-modal approaches learn the correspondence between different modalities of the input. Different modalities can accelerate compact learning in a single target modality by exploiting cross-modality structure. OpenL3 (§4, Cramer et al. (2019)) is a broad-domain audio model trained on the correspondence between audio and video. Contrastive Language-Image Pre-training (CLIP, Radford et al. (2021)) learns a model from 400 M image-text pairs, and was successfully applied on zero-shot tasks. Wang and van den Oord (2021) contrastively induce audio representations from waveforms (1-D audio) and spectrograms, and Wang et al. (2021b) extend that to include correspondence with video frames.

Because pretraining large-scale models requires large quantities of data and can be computationally expensive, another research direction has been on distilling information from existing models that were trained on another modality for which more data are available. Aytar et al. (2016) train SoundNET which distills audio representations from a pre-trained image classification model trained on large image datasets such as ImageNet (Deng et al., 2009). Wu et al. (2022a) (§4) distill an audio representation (Wav2CLIP) from a large text-image model (CLIP) using video data to link the visual and audio modalities.

Using and evaluating representations Representation models can be used in downstream tasks with full fine-tuning; but the naïve approach is simply to treat intermediate pre-trained model outputs as frozen embeddings, and this nonetheless provides a stark improvement over using raw features (Turian et al., 2010). Broad-scale evaluation of learned representations has been done in other ML domains, in NLP, for example: GLUE (Wang et al., 2019b), the harder SuperGLUE (Wang et al., 2019a), and ERASER (DeYoung et al., 2020). Vision includes the FAIR self-supervision benchmark (Goyal et al., 2019) and VTAB (Zhai et al., 2019).

3. HEAR: Holistic Evaluation of Audio Representations

A strong general-purpose audio representation should be as holistic as the human ear. The goal of the HEAR competition is to evaluate audio representations across a variety of everyday domains, audio phenomena, with tasks that involve short and long time spans, sometimes with few labeled instances. Formal rules are provided on the HEAR website.²

3.1. Related audio shared tasks

Historical audio shared tasks, such as those from MIREX (Downie et al., 2014), DCASE (Mesaros et al., 2017), and INTERSPEECH ComParE (Schuller et al., 2013) have improved the community’s understanding of audio modeling substantially. However, the bespoke nature of these tasks is a double-edged sword, requiring substantial custom tooling both by the challenge organizers and participants. More recent audio shared tasks focus on reusability and generic task APIs. SUPERB (Yang et al., 2021) focuses on a broad spectrum of speech tasks, and includes downstream evaluation ranging from simple classification to LSTM-based sequence modeling. Although the Speech Commands v2 task is shared with HEAR, the other downstream tasks in SUPERB mainly deal with speech processing applications, including speech recognition, speaker verification, keyword spotting, etc., and these two evaluation activities are complementary to each other. The NOSS (Non-Semantic Speech Benchmark) (Shor et al. (2020)) comprises 6 paralinguistic tasks. Two tasks are shared with HEAR: CREMA-D and Speech Commands v2. Unfortunately, SAVEE and DementiaBank require filling out a request form, and VoxCeleb requires scraping YouTube. HARES (Holistic Audio Representation Evaluation Suite)—not to be confused with our HEAR challenge—is concurrently published work (Wang et al., 2021c). HARES comprises 12 well-known downstream tasks including—like HEAR—ESC-50, Speech Commands v2, and an NSynth Pitch task, benchmarked on 13 models. Where HARES differs from HEAR includes: a) HARES tasks are well-known benchmarks, whereas HEAR is a mix of well-known and novel benchmarks, b) HARES includes no few-shot tasks, all tasks have $\geq 2K$ samples, c) HARES results currently include no external submissions, d) evaluation code and dataset links are not provided and e) two of the tasks (AudioSet and VoxCeleb) tasks involve scraping YouTube. Datasets based upon YouTube require specialized code and lack reproducibility because videos are removed unpredictably (Cramer et al., 2019). These generic audio evaluation suites, including our HEAR challenge, intend to make it easy to evaluate existing models on novel tasks, at the expense of possible SOTA performance.

3.2. Evaluation methodology

Wrapping existing models into the HEAR API requires roughly 75 lines of code, much of which is boilerplate. New HEAR tasks can be run with no code changes. HEAR 2021 includes two types of tasks: 1) Scene-based: Multi-class or multi-label classification of an entire audio clip; 2) Timestamp-based: Sound event detection/transcription, which involves detecting when exactly sound events occur over time by providing a start time, end time, and label for each sound event. In both cases, the audio representation is frozen and used as the input feature vector to a shallow downstream MLP classifier, with no fine-tuning.

2. <https://neuralaudio.ai/hear2021-rules.html>

Fine-tuning improves downstream performance (Baevski et al. (2020); Shor et al. (2020)), but increases training time. Crucially, the use of frozen embeddings means that HEAR downstream evaluation code can be maintained solely in PyTorch, regardless of whether the embedding model was in TensorFlow or PyTorch.³

A timestamp-based task can be simplified to a frame-based sequence-labeling task of the audio at regular intervals (Kelz et al., 2016), and we use a common postprocessing step to compose predictions from multiple timesteps and extract discrete labeled events with start and ends times (Mesaros et al., 2016). Framewise accuracy (the decomposed multilabel prediction, computed at regular timesteps) does not always correlate well with the perceptual quality of event-onset FMS (Hawthorne et al., 2018) because they ignore the interplay between the frame representations and more sophisticated downstream inference (Cheuk et al., 2021). See Section B for details on the downstream training regime.

3.3. Evaluation tasks

The following are the evaluation tasks for HEAR 2021. For simplicity and reproducibility, we have preprocessed each relevant datasets to all commonly used sample rates (16000, 22050, 32000, 44100), fixed the length of the audio clips, predefined training splits, and packaged each dataset in a self-explanatory common format with human-readable metadata. They all have open licenses (some of which permit commercial use), with the exception of the GTZAN corpora which are widely used but of unknown license status. We encourage the community to use our datasets, even if they do not follow all or any the HEAR 2021 rules, but encourage deviations from the HEAR 2021 rules to be described.

Open tasks were released early in the challenge, to encourage participation and to allow participants to debug and refine their submissions: Speech Commands v2 (full and 5h versions), NSynth Pitch (50h and 5h versions), and DCASE 2016 Task 2. Tasks are summarized in Table 1 described with more detail in Table 2 and Section A.

4. Models evaluated

Evaluated models are described below. Table 3 summarizes model properties. HEAR began with three strong baseline models (§4.1), each pretrained on a different audio domain. We report on 13 external teams’ submissions to HEAR 2021 (§4.2).

4.1. Baseline models

wav2vec2 wav2vec2 (1-D CNN and positional transformer) (Baevski et al., 2020). Self-supervised pretraining on 100K hours of speech from VoxPopuli (Wang et al., 2021a).

CREPE 1-D CNN. Supervised pretraining of pitch-tracking on 16 hours of synthesized music. (Kim et al., 2018b)

3. We initially believed that imposing a restriction that all submitted models must be TensorFlow 2.x or PyTorch and pip3-installable would facilitate easy orchestration of model testing. However, models submitted with competing TensorFlow, CUDA, CuDNN, and pypi dependencies lead us to suggest that future ML challenge organizers standardize on the latest stable microversion of all deep learning packages.

Table 1: HEAR tasks.

Speech Commands (version 2), 5h and full Spoken commands classification.

NSynth Pitch, 5h and 50h Pitch classification of synthesized sounds.

DCASE 2016 Task 2 Office sound event detection in synthesized scenes.

Beehive States Binary classification of normal vs. queen-less beehives.

Beijing Opera Percussion Classification of six Beijing Opera percussion instruments.

CREMA-D Speech emotion recognition.

ESC-50 Environmental sound classification.

FSD50K Broad-domain audio multi-labeling.

Gunshot Triangulation Identify location of microphone recording a gunshot, using classification.

GTZAN Genre Music genre classification.

GTZAN Music Speech Classification of audio into music or speech.

LibriCount Multiclass speaker count identification.

MAESTRO 5h Music transcription.

Mridingham Stroke and Mridingham Tonic Non-Western pitched percussion. Classification of stroke or tonic.

Vocal Imitations Match a vocal imitation to the type of sound imitated, using classification.

VoxLingua107 Top 10 Spoken language identification.

OpenL3 2-D CNN. Multi-modal contrastive self-supervised pretraining of audio/video correspondence on 6K hours of AudioSet broad-domain YouTube content. ([Cramer et al. \(2019\)](#), earlier [Arandjelovic and Zisserman \(2017\)](#)) HEAR implementation by Jon Nordby.

4.2. Submitted models

AMAAI Lab SUTD wav2vec2+DDSP An ensemble of wav2vec2 ([Baevski et al., 2020](#)) and two DDSP encoders ([Engel et al., 2020](#)). The wav2vec2 model is pretrained on the Librispeech ([Panayotov et al., 2015](#)) and MAESTRO ([Hawthorne et al., 2019](#)) datasets. One DDSP encoder is CREPE, the other is a non-pretrained loudness encoder.

AMAAI wav2vec2 music+speech wav2vec2 model ([Baevski et al., 2020](#)). Pretrained on Librispeech ([Panayotov et al., 2015](#)) and MAESTRO ([Hawthorne et al., 2019](#)).

Audioshake UDONS ViT Vision transformer (ViT, [Kolesnikov et al. \(2021\)](#)). Pre-trained on 360h of Librispeech to predict the correct permutation ([Noroozi and Favaro, 2016](#); [Carr et al., 2021](#)) of up to 5 patches of mel-spectrogram input, shuffled in time.

CP-JKU PaSST base, base2level, base2levelmel Patchout fast (2-D) spectrogram transformer (PaSST, [Koutini et al. \(2021\)](#)). Initialized from a ImageNet vision transformer model, and further pretrained on 10s audio from AudioSet to perform supervised tagging. base2level concatenates a longer window (160 ms and 800ms) for timestamp embeddings. base2levelmel additionally concatenates the raw melspectrogram as well.

CVSSP (University of Surrey) PANNs 2-D CNN14. Pretrained on AudioSet with supervision (Kong et al., 2020).

Descript/MARL Wav2CLIP 2-D ResNet18. Pretrained multimodally using contrastive learning on the 600h VGGSound corpus (Chen et al., 2020a) (without supervised labels) to distill the Contrastive Language-Image Pre-training (CLIP, Radford et al. (2021)) language and image model to a corresponding audio embedding (Wu et al., 2022a).

IUT-CSE kwmlp and audiomlp Sequentially stacked gated MLP model (Liu et al., 2021), taking (2-D) MFFCs as input. kwmlp (Morshed et al., 2022) is pretrained with supervision on Speech Commands v2. audiomlp is pretrained with supervision on HEAR open task datasets: Speech Commands v2, DCASE 2016 Task 2, and NSynth Pitch.

Kuroyanagi hearline 2-D conformer model. Pretraining unknown.

Logitech AI SERAB BYOL-S 2-D CNN. Self-supervised pretraining using the BYOL self-supervised approach (Grill et al., 2020) adapted to audio (BYOL-A, Niizumi et al. (2021)), pretrained on the speech subset of AudioSet (Elbanna et al., 2022).

NTU-GURA (fusion) avg/cat hubert/wav2vec2/crepe Three models (HuBERT Hsu et al. (2021), wav2vec2, CREPE) combined in a variety of ways: averaged or concatenated (Wu et al., 2022b). Fusion of multiple model layers was optionally included. fusion_cat_xwc_time is a variation of fusion_cat_xwc with a different approach to matching timestamps when concatenating different models' emmbeddings.

RedRice/Xiaomi EfficientNet-B2 2-D EfficientNet-B2 (Tan and Le, 2019). Pretrained on supervised AudioSet tags. Instead of global averaging pooling, decision-level pooling is used. Timestamp embeddings are smeared scene embeddings.

Soundsensing YAMNet 2-D MobileNet (Howard et al., 2017). Pretrained to tag AudioSet.

Stellenbosch LSL Audio DBERT 1-D CNN encoder and modified BERT transformer. Pretrained as the discriminator with a GAN objective, using the clustering model as the generator, on 960 hours of Librispeech (Panayotov et al., 2015). Embeddings are taken from layer 16 of 24 by default.

5. Results and Discussion

In Figure 1 we present the primary score of submitted models on each HEAR 2021 task. By default, evaluation uses a deterministic seed, for reproducibility. Nonetheless, scores are stable across our evaluation, with a median 95% confidence interval of 2.5e-3 when seeding of model weights and hyperparameter grid points is selected non-deterministically. Shor et al. (2020); Wu et al. (2022a) present scores for some of the the same models and tasks. HEAR reported scores are similar but not identical, due to downstream training differences.

To display model similarity at a glance, we present t-SNE visualizations of normalized scores by task (Figure 2(a)) and by model (Figure 2(b)). We also show correlation tables for tasks (Figure 3) and models (Figure 4) to give greater insight into model and task similarity, in similar spirit to the confusion matrices of Wu et al. (2022a). Zhai et al. (2019)

compare a variety of aggregation techniques for evaluating cross-task model performance, and find that they are all highly correlated, settling upon simple mean top-1. [Gosiewska et al. \(2020\)](#) proposes an ELO-like meta-score for cross-task model performance, similar to a chess rating. Although it is tempting to give a single score for every model, we believe that would strip out important nuances shown in the full score table ([DeYoung et al., 2020](#)).

For these summary figures, we normalize each model/task score. Normalized scores allow us to compare models and tasks against each other, under the assumption each task is equally weighted. The normalization procedure is as follows: 1) For each task, we standardize the scores to zero mean and unit variance. Unlike transforming tasks to ranks, we assume that the scale of intra-task scores is important. 2) The standardized scores are Winsorized (clamped) to have variance within $[-1, +1]$. By limiting the importance of extremely high or low scores on a single task, this approach allows for better inter-task comparison.

In the following paragraphs, we describe a few interesting patterns and trends in the submitted models. Many evaluated models use the last layer as the representation. It is known that non-final layers and/or fusing various layers might capture more information ([Shor et al., 2020](#); [Baeviski, 2020](#); [Yang et al., 2021](#)). Intermediate layers often model audio phenomena that are not necessary for the final loss. NTU-GURA’s ablation studies support that, as evidence by the relative performance of their different models. For conciseness, we use the term “strong speech models” to refer to NTU-GURA’s fused models that include pretrained speech models.

Pitch tasks NSynth pitch and Maestro tasks have similar results, and models that include CREPE embeddings ([Kim et al., 2018b](#)) perform best. This makes sense as these tasks require modeling of pitch, which CREPE was specifically trained for, while many other representations focus on discriminating between semantic objects (e.g., cat vs dog or guitar vs piano) but are pitch agnostic. Interestingly, models trained for semantic discrimination (e.g., via AudioSet) and speech models do nonetheless represent pitch to some degree, as evidenced by the decent performance of OpenL3 and wav2vec2 on these tasks.

Broad Domain Semantic-Object Tagging FSD50K and ESC-50 semantic-object tagging results are strongly correlated, as well as—perhaps surprisingly—GTZAN genre tagging. The models that perform the best on this group are the ones pretrained on the AudioSet semantic-object tagging task. What we glean from this large-scale survey of diverse models is that results on ESC-50 and GTZAN genre tagging are strongly predictive of results on the more nuanced FSD50K task, despite being an order of magnitude smaller and not using the corrected GTZAN artist-conditional splits from ([Sturm, 2013](#)), suggesting faster inroads for research iteration. One valuable point-based contribution of HEAR 2021 is that the CP-JKU PaSST base model achieves a new state-of-the-art on FSD50K despite no fine-tuning, a mean average precision (mAP) of 0.640 on FSD50K, compared to the recent literature ([Gong et al., 2021b](#); [Wu et al., 2022a](#); [Fonseca et al., 2021a](#)).

Vocals FSD50K scores are also similar to those of Vocal Imitations and LibriCount. This is perhaps because Vocal Imitations comprises broad non-semantic vocalizations and LibriCount involves detecting multiple simultaneous audio events. The strong speech and PaSST models do the best on Vocal Imitations. On LibriCount, SERAB BYOL-S does the best as a non-semantic speech model, with decent performance from strong speech models.

HEAR 2021: HOLISTIC EVALUATION OF AUDIO REPRESENTATIONS

model	task																			
	Beehive	Beijing Opera	CREMA-D	DCASE 2016	ESC-50	FSD50k	GTZAN Genre	GTZAN Music/Speech	Gunshot	Libricount	Maestro 5h	Mridangam Stroke	Mridangam Tonic	NSynth Pitch 50h	NSynth Pitch 5h	Speech commands 5h	Speech commands full	Vocal Imitation	VoxLingua107 top 10	
GURA Fuse Cat H+w+C	.966	.747	.826	.734	.420	.805	.928	.935	.697	.441	.972	.923	.885	.846	.961	.968	.197	.720		
GURA Fuse Cat H+w+C (t)	.962	.743	.826	.653	.374	.760	.944	.905	.659	.441	.975	.924	.891	.854	.951	.968	.215	.629		
GURA Fuse Hubert	.949	.752	.826	.743	.413	.796	.936	.929	.683	.166	.974	.909	.688	.382	.947	.957	.185	.714		
GURA Fuse wav2vec2	.945	.692	.798	.695	.403	.793	.953	.967	.653	.111	.962	.838	.606	.330	.957	.969	.174	.706		
Logitech SERAB BYOL-S	.549	.953	.657	.642	.805	.509	.837	.938	.857	.785	.008	.973	.928	.712	.396	.914	.948	.160	.458	
GURA Cat H+w+C	.936	.639	.681	.511	.314	.722	.961	.881	.639	.469	.938	.859	.897	.866	.927	.943	.111	.460		
OpenL3	.604	.975	.550	.833	.751	.447	.879	.969	.949	.641	.017	.967	.937	.731	.560	.680	.763	.078	.331	
CP-JKU PaSST 2lvi+mel	.966	.610	.925	.947	.558	.883	.977	.940	.660		.965	.819	.541	.256	.681	.639	.182	.259		
CP-JKU PaSST 2lvi	.966	.610	.913	.947	.537	.883	.977	.940	.660		.965	.819	.541	.256	.681	.639	.182	.259		
CP-JKU PaSST base	.966	.610	.788	.947	.640	.883	.977	.940	.660		.965	.819	.541	.256	.681	.639	.182	.259		
GURA Hubert	.945	.690	.584	.603	.314	.735	.913	.932	.646	.007	.953	.850	.429	.184	.953	.954	.154	.637		
GURA Avg H+w+C	.945	.547	.624	.450	.264	.706	.937	.857	.631	.460	.914	.837	.896	.862	.822	.881	.084	.321		
GURA Cat Hubert+wav2vec2	.936	.698	.452	.586	.323	.734	.936	.869	.627	.003	.951	.802	.428	.198	.936	.929	.166	.706		
wav2vec2	.907	.656	.663	.561	.116	.780	.946	.848	.692	.033	.943	.828	.653	.402	.838	.879	.080	.493		
RedRice EfficientNet-B2	.533	.953	.575	.790	.935	.607	.878	.968	.878	.651	.000	.949	.843	.391	.168	.573	.676	.138	.255	
GURA Avg Hubert+CREPE	.932	.540	.610	.437	.253	.698	.946	.845	.622	.462	.917	.831	.897	.878	.737	.823	.079	.293		
GURA Avg Hubert+wav2vec2	.941	.699	.417	.587	.318	.746	.928	.810	.623	.004	.946	.799	.415	.198	.907	.952	.165	.690		
GURA Cat wav2vec2+CREPE	.920	.460	.585	.343	.209	.681	.938	.833	.569	.463	.898	.823	.899	.868	.885	.919	.076	.310		
CVSSP PANNS	.446	.911	.555	.000	.909		.860	.992	.798	.652	.000	.939	.824	.301	.148	.560	.618	.127	.244	
IUT-CSE MLP (keyword)	.760	.911	.424	.518	.367	.187	.554	.889	.932	.451	.065	.969	.942	.605	.440	.976	.978	.056	.181	
Stellenbosch LSL DBERT	.697	.919	.522	.246	.532	.263	.674	.969	.810	.584	.000	.835	.685	.737	.524	.566	.630	.161	.246	
Descript/MARL Wav2CLIP	.770	.936	.512	.000	.759	.362	.748	.946	.929	.528	.000	.947	.829	.439	.230	.316	.347	.083	.192	
Audioshake UDONS ViT	.878	.928	.441	.668	.401		.681	.899	.866	.488	.239	.880	.730	.795	.692	.479	.531	.068	.224	
Soundsensing YAMNet	.466	.941	.453	.008	.838		.847	.969	.732	.653	.000			.321	.158	.289	.410	.085	.202	
CREPE	.593	.928	.383	.504	.300	.159	.645	.929	.863	.499	.401	.898	.824	.900	.870	.180	.211	.051	.142	
Kuroyanagi hearline		.928	.480		.424	.178	.654	.931	.518	.498	.002	.909	.783	.589	.394	.478	.580	.040	.180	
AMAAI w2v2 music+spch		.826	.391	.510	.511	.198	.605	.907	.857	.432	.003	.759	.415	.176	.182	.744	.523	.064	.169	
IUT-CSE MLP (audio)	.535	.728	.420	.052	.161	.086	.408	.774	.786	.366	.024	.893	.778	.608	.386	.392	.535	.038	.129	
AMAAI Lab w2v2+DDSP		.635		.507	.280	.268			.583	.510	.000	.653	.357	.303	.172	.487	.490	.072	.106	

Figure 1: Primary score of submitted models on each HEAR 2021 task. Normalized scores are used to show the heat-value of each cell. Missing cells indicate that the model did not successfully complete the task (exhausting GPU memory or exceeding 24 hours downstream training time).

Speech As we move into the speech domain, LibriCount and Vocal Imitations have the most similarity to CREMA-D emotion detection, which then is most similar to VoxLingua107 Top 10 language identification, which in turn is correlated with Speech Commands, following a trend from “environmental” to paralinguistic to semantic. The strong speech models do the best on these tasks.

What is most interesting about our diverse survey of 29 models \times 19 tasks is, perhaps, the most difficult to explain results: tasks that defy neat categorization suggest the fragile, unpredictable boundaries of existing models. DCASE 2016 Task 2 seems *a priori* similar to FSD50K and ESC-50, but not in practice. Vocal Imitations are human-depictions of all kinds of sounds. Gunshot Triangulation is an extremely low-resource task with only 88 instances. Beijing Opera and Mridingham Stroke and Tonic are non-Western music tasks. For these tasks, our contribution is a negative result: we have no simple story or obvious pretraining data to attack them. Robust generalization of >10 -billion-parameter models from NLP (Brown et al., 2020) and vision (Goyal et al., 2022) suggest one path forward.

6. Conclusion

General-purpose models that transfer to few-shot and zero-shot scenarios are highly desirable. The audio community has followed the NLP and vision communities in using increasingly sophisticated representation learning approaches. The HEAR challenge allows the audio community also to follow the trend of broad-scale reproducible evaluation.

HEAR 2021 is about openness. The datasets and the submissions are as open as possible. All HEAR 2021 datasets are preprocessed to a common format with standard splits, and distributed as tarfiles. This alleviates the risk of dataset rot common in YouTube scraping, and the difficulty of acquiring data locked behind closed-access request forms. All HEAR 2021 submissions have code that is Apache 2.0 compatible, models that are CC-Attribution compatible, and follow a common API, so switching between them requires a single line of code. Evaluation code, submitted models, and datasets are key contributions of HEAR 2021, available at <https://neuralaudio.ai/hear.html>.

Twenty-nine models were evaluated on 19 diverse downstream tasks, spanning speech, environmental sounds, and music, and datasets that don’t fit neatly into any rubric, as well as datasets that span the boundaries of multiple audio domains. This large standardized set of tasks and models pave the way for comprehensive and reproducible evaluation, enabling previously impossible longitudinal studies. We are eager to help onboard new tasks into the HEAR benchmark suite, particularly unusual and/or few-shot audio tasks. The largest-scale HEAR scene-embedding tasks and the CPU-gated evaluation of timestamp-embedding tasks were the most difficult tasks to run, sometimes requiring 24 hours for downstream evaluation of a single model-task pair on an A100 GPU, despite no fine-tuning.

Before an evaluation like HEAR, it would be easy for the community to suggest which audio tasks are predictably hard: large-scale, well-defined datasets with no more low-hanging fruit that are known to be difficult to hill-climb. Our contribution—the existence and easy accessibility of HEAR datasets, models, and evaluation code—allows the community to probe what we don’t know. And the central question posed by HEAR 2021 remains open: Can one single general-purpose audio representation perform as holistically as the human ear? If one does, then there is clearly more work to be done towards achieving it.

Acknowledgments

HEAR 2021 was sponsored by Google, and competition evaluation was performed on Google Cloud Platform.

References

- Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, and Björn Schuller. Snore Sound Classification Using Image-based Deep Spectrum Features. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3512–3516, Stockholm, Sweden, 2017.
- Akshay Anantapadmanabhan, Ashwin Bellur, and Hema A Murthy. Modal analysis and transcription of strokes of the Mridangam using non-negative matrix factorization. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–185, 2013.
- Joakim Andén and Stéphane Mallat. Deep Scattering Spectrum. *IEEE Transactions on Signal Processing*, 62(16):4114–4128, 2014.
- Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 609–617, 2017.
- Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, volume 29, 2016.
- Alexei Baevski. Need to extract feature from best_checkpoints of wav2vec2.0 model – Github comment. <https://github.com/pytorch/fairseq/issues/2495#issuecomment-701619177>, 2020.
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, 2020.
- Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *to appear at International Conference on Learning Representations (ICLR)*, 2022. URL <http://arxiv.org/abs/2105.04906>.
- Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A Neural Probabilistic Language Model. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, 2001.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen,

- Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.
- Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.
- Andrew N Carr, Quentin Berthet, Mathieu Blondel, Olivier Teboul, and Neil Zeghidour. Self-Supervised Learning of Audio Representations From Permutations With Differentiable Ranking. *IEEE Signal Processing Letters*, 28:708–712, 2021.
- Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A Large-scale Audio-Visual Dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725, 2020a.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, 2020b.
- Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Kin Wai Cheuk, Yin-Jyun Luo, Emmanouil Benetos, and Dorian Herremans. Revisiting the Onsets and Frames Model with Additive Attention. In *International Joint Conference on Neural Networks, (IJCNN)*, pages 1–8, 2021.
- Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 539–546, 2005.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.
- Seth Cooper and Steven Shaw. Gunshots recorded in an open field using iPod Touch devices, 2020. URL <https://doi.org/10.5061/dryad.wm37pvmkc>.
- Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856, 2019.
- Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4443–4458, 2020.
- J Stephen Downie, Xiao Hu, Jin Ha Lee, Kahyun Choi, Sally Jo Cunningham, and Yun Hao. Ten years of MIREX: reflections, challenges and opportunities. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pages 657–662. ISMIR, 2014.
- Gasser Elbanna, Neil Scheidwasser-Clow, Mikolaj Kegler, Pierre Beckmann, and Milos Cernak. Byol-s: Learning self-supervised speech representations by bootstrapping. In Joseph Turian, Björn W. Schuller, Dorien Herremans, Katrin Kirchhoff, Paola Garcia Perera, and Philippe Esling, editors, *Proceedings of HEAR 2021: Holistic Evaluation of Audio Representations*, volume 166 of *Proceedings of Machine Learning Research*. PMLR, 2022. In submission.
- Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable Digital Signal Processing. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.
- Jesse H. Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1068–1077, 2017.
- Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an Open Dataset of Human-Labeled Sound Events. *CoRR arXiv*, 2020. URL <http://arxiv.org/abs/2010.00475>.
- Eduardo Fonseca, Andres Ferraro, and Xavier Serra. Improving sound event classification by increasing shift invariance in convolutional neural networks. *CoRR arXiv*, 2021a. URL <https://arxiv.org/abs/2107.00623>.
- Eduardo Fonseca, Diego Ortego, Kevin McGuinness, Noel E O’Connor, and Xavier Serra. Unsupervised contrastive learning of sound event representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 371–375, 2021b.

- Jort F Gemmeke, Daniel P W Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.
- Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterton, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 2010. PMLR.
- Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021a.
- Yuan Gong, Yu-An Chung, and James Glass. PSLA: Improving Audio Tagging With Pre-training, Sampling, Labeling, and Aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3292–3306, 2021b.
- Alicja Gosiewska, Katarzyna Woznica, and Przemyslaw Biecek. Interpretable Meta-Measure for Model Performance. *CoRR arXiv*, 2020. URL <http://arxiv.org/abs/2006.02293>.
- Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and Benchmarking Self-Supervised Visual Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6390–6399, 2019. URL <https://github.com/facebookresearch/>.
- Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncured images without supervision. *CoRR arXiv*, 2022. URL <https://arxiv.org/abs/2202.08360>.
- Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap Your Own Latent – A New Approach to Self-Supervised Learning. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.
- Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. Onsets and Frames: Dual-Objective Piano Transcription. In Emilia Gómez, Xiao Hu, Eric Humphrey, and Emmanouil Benetos, editors, *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pages 50–57, 2018.
- Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling Factorized Piano

- Music Modeling and Generation with the MAESTRO Dataset. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.
- Shawn Hershey, Sourish Chaudhuri, Daniel P W Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, Malcolm Slaney, Ron J Weiss, and Kevin Wilson. CNN Architectures for Large-Scale Audio Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.
- Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR arXiv*, 2017. URL <http://arxiv.org/abs/1704.04861>.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.
- Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning (ICML)*, pages 2410–2419. PMLR, 2018.
- Rainer Kelz, Matthias Dorfer, Filip Korzeniowski, Sebastian Böck, Andreas Arzt, and Gerhard Widmer. On the Potential of Simple Framewise Approaches to Piano Transcription. In Michael I. Mandel, Johanna Devaney, Douglas Turnbull, and George Tzanetakis, editors, *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pages 475–481, 2016.
- Bongjun Kim, Madhav Ghei, Bryan Pardo, and Zhiyao Duan. Vocal imitation set: a dataset of vocally imitated sound events using the AudioSet ontology. In Mark D. Plumbley, Christian Kroos, Juan P. Bello, Gaël Richard, Daniel P. W. Ellis, and Annamaria Mesaros, editors, *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, volume 1, pages 148–152, 2018a.
- Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A Convolutional Representation for Pitch Estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, 2018b.
- Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
- Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition.

- IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.
- Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient Training of Audio Transformers with Patchout. *CoRR arXiv*, 2021. URL <http://arxiv.org/abs/2110.05069>.
- Kushal Lakhotia, Eugene Khartitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On Generative Spoken Language Modeling from Raw Audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.
- Hanxiao Liu, Zihang Dai, David So, and Quoc Le. Pay Attention to MLPs. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Beth Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Society for Music Information Retrieval Conference (ISMIR)*, Plymouth, Mass, USA, 2000.
- Soroush Merhi, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron C. Courville, and Yoshua Bengio. SampleRNN: An Unconditional End-to-End Neural Audio Generation Model. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.
- Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6), 2016.
- Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, and Mark D Plumbley. Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):379–393, 2017.
- Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, and Mark D Plumbley. Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):379–393, 2018.
- Mashrur M Morshed, Ahmad Omar Ahsan, Hasan Mahmud, and Md. Kamrul Hasan. Learning Audio Representations With MLPs. In Joseph Turian, Björn W. Schuller, Dorian Herremans, Katrin Kirchhoff, Paola Garcia Perera, and Philippe Esling, editors, *Proceedings of HEAR 2021: Holistic Evaluation of Audio Representations*, volume 166 of *Proceedings of Machine Learning Research*. PMLR, 2022. In submission to HEAR-2021.
- Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.
- Inês Nolasco, Alessandro Terenzi, Stefania Cecchi, Simone Orcioni, Helen L Bear, and Emmanouil Benetos. Audio-based Identification of Beehive States. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8256–8260, 2019.

- Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, volume 9910 of *Lecture Notes in Computer Science*, pages 69–84. Springer, 2016.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- Santiago Pascual, Mirco Ravanelli, Joan Serra, Antonio Bonafonte, and Yoshua Bengio. Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks. In Gernot Kubin and Zdravko Kacic, editors, *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 161–165, 2019.
- Karol J Piczak. ESC: Dataset for Environmental Sound Classification. In Xiaofang Zhou, Alan F. Smeaton, Qi Tian, Dick C. A. Bulterman, Heng Tao Shen, Ketan Mayer-Patel, and Shuicheng Yan, editors, *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference (SIGMM)*, pages 1015–1018, 2015.
- Jordi Pons and Xavier Serra. Randomly Weighted CNNs for (Music) Audio Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340, 2019.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 Classifiers Generalize to CIFAR-10? *CoRR arXiv*, 2018. URL <http://arxiv.org/abs/1806.00451>.
- Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879, 2021.
- Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3): 279–283, 2017.
- Andrew M Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y Ng. On Random Weights and Unsupervised Feature Learning. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 1089–1096, 2011.

- Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised Pre-Training for Speech Recognition. In Gernot Kubin and Zdravko Kacic, editors, *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3465–3469, 2019.
- Christian Schörkhuber and Anssi Klapuri. Constant-Q transform toolbox for music processing. In *Proceedings of the 7th Sound and Music Computing Conference*, pages 3–64, 2010.
- Björn Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2013 Computational Paralinguistics Challenge – A Brief Review. *Speech and Language Processing Technical Committee (SLTC) Newsletter*, 2013.
- Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Felix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. Towards Learning a Universal Non-Semantic Representation of Speech. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 140–144, 2020.
- Fabian-Robert Stöter, Soumitro Chakrabarty, Bernd Edler, and Emanuël A P Habets. Classification vs. Regression in Supervised Learning for Single Channel Speaker Count Estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 436–440, 2018a.
- Fabian-Robert Stöter, Soumitro Chakrabarty, Emanuël Habets, and Bernd Edler. LibriCount, a dataset for speaker count estimation, 2018b. URL <https://doi.org/10.5281/zenodo.1216072>.
- Bob L. Sturm. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *CoRR arXiv*, 2013. URL <http://arxiv.org/abs/1306.1461>.
- Mingxing Tan and Quoc V Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 6105–6114, 2019.
- Mi Tian, Ajay Srinivasamurthy, Mark Sandler, and Xavier Serra. A study of instrument-wise onset detection in Beijing Opera percussion ensembles. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2159–2163, 2014. doi: 10.1109/ICASSP.2014.6853981. URL <http://dx.doi.org/10.1109/ICASSP.2014.6853981>.
- Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What Makes for Good Views for Contrastive Learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.
- Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Peter Steiner, Daniel Keysers, Jakob Uszkoreit, Mario

- Lucic, and Alexey Dosovitskiy. Pay Attention to MLPs. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- George Trigeorgis, Fabien Ringeval, Raymond Brückner, Erik Marchi, Mihalis Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network. In *Proceedings of the 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5200–5204, Shanghai, China, 2016.
- Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word Representations: A Simple and General Method for Semi-Supervised Learning. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 384–394, 2010.
- George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.
- Jörgen Valk and Tanel Alumäe. VOXLINGUA107: A Dataset for Spoken Language Recognition. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658, 2021.
- Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *Proceedings of the 9th ISCA Speech Synthesis Workshop*, page 125, 2016.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 3261–3275, 2019a.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019b.
- Changhan Wang, Morgane Rivi re, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 993–1003, 2021a.
- Luyu Wang and Aaron van den Oord. Multi-Format Contrastive Learning of Audio Representations. *CoRR arXiv*, 2021. URL <http://arxiv.org/abs/2103.06508>.

- Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord. Multimodal Self-Supervised Learning of General Audio Representations. *CoRR arXiv*, 2021b. URL <http://arxiv.org/abs/2104.12807>.
- Luyu Wang, Pauline Luc, Yan Wu, Adria Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, Joao Carreira, and Aaron van den Oord. Towards Learning Universal Audio Representations. *CoRR arXiv*, 2021c. URL <http://arxiv.org/abs/2111.12124>.
- Pete Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *CoRR arXiv*, 2018. URL <http://arxiv.org/abs/1804.03209>.
- Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2CLIP: Learning robust audio representations from CLIP. In *to appear in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022a.
- Tung-Yu Wu, Chen-An Li, Tsu-Yuan Hsu, and Tzu-Han Lin. The ability of self-supervised speech models for audio representations. In Joseph Turian, Björn W. Schuller, Dorien Herremans, Katrin Kirchhoff, Paola Garcia Perera, and Philippe Esling, editors, *Proceedings of HEAR 2021: Holistic Evaluation of Audio Representations*, volume 166 of *Proceedings of Machine Learning Research*. PMLR, 2022b. In submission.
- Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-Tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-Yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1194–1198, 2021.
- Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320, 2021.
- Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. LEAF: A Learnable Frontend for Audio Classification. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
- Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark. *CoRR arXiv*, 2019. URL <http://arxiv.org/abs/1910.04867>.

Table 2: Summary of the 19 evaluation tasks for HEAR 2021. Includes the embedding type (timestamp (T) or scene (S)), the predictor type (multiclass (C) or multilabel (L)), the split method used during downstream evaluation (train/validation/test (TVT) or K -fold), the duration of clips in seconds, the total number of clips for each task, the primary evaluation metric, and whether or not the task is novel. Novel tasks are not comparable to the literature. For all tasks except FSD50k, clips were standardized to one length using padding or trimming, typically the 95th percentile length in the original corpus.

Task Name	Embed Type	Predictor Type	Split Method	Duration (seconds)	# clips	Evaluation Metric	Novel
Open Tasks							
DCASE 2016 Task 2	T	L	TVT	120.0	72	Onset FMS	✓
NSynth Pitch 5hr	S	C	TVT	4.0	5000	Pitch Acc.	✓
NSynth Pitch 50hr	S	C	TVT	4.0	49060	Pitch Acc.	✓
Speech Commands 5hr	S	C	TVT	1.0	22890	Accuracy	✓
Speech Commands Full	S	C	TVT	1.0	100503	Accuracy	
Secret Tasks							
Beehive States	S	C	TVT	600.0	576	AUCROC	
Beijing Opera Percussion	S	C	5-fold	4.77	236	Accuracy	✓
CREMA-D	S	C	5-fold	5.0	7438	Accuracy	
ESC-50	S	C	5-fold	5.0	2000	Accuracy	
FSD50K	S	L	TVT	0.3 - 30.0	51185	mAP	
Gunshot Triangulation	S	C	7-fold	1.5	88	Accuracy	✓
GTZAN Genre	S	C	10-fold	30.0	1000	Accuracy	
GTZAN Music Speech	S	C	10-fold	30.0	128	Accuracy	
LibriCount	S	C	5-fold	5.0	5720	Accuracy	
MAESTRO 5hr	T	L	5-fold	120.0	185	Onset FMS	✓
Mridangam Stroke	S	C	5-fold	0.81	6977	Accuracy	✓
Mridangam Tonic	S	C	5-fold	0.81	6977	Accuracy	✓
Vocal Imitations	S	C	3-fold	11.26	5601	mAP	✓
VoxLingua107 Top10	S	C	5-fold	18.64	972	Accuracy	✓

Appendix A. Evaluation Tasks

Our 19 tasks were derived from 16 datasets, as described in more detail below. Tasks described as “novel” are not comparable to the literature. A summary of task statistics is available in Table 2.

Speech Commands (version 2), 5h and full Classification of known spoken commands, with additional categories for silence and unknown commands (Warden, 2018). As per the literature, models are evaluated by prediction accuracy. For this challenge we also provide a 5-hour subset of the training data. We use the predefined train and test split, and note that the test data has a different distribution of labels from the training data.

NSynth Pitch, 5h and 50h NSynth Pitch is a novel multiclass classification problem. The goal of this task is to classify instrumental sounds from the NSynth Dataset (Engel et al., 2017) into one of 88 pitches. Results for this task are measured by pitch accuracy, as well as chroma accuracy. Chroma accuracy considers only the pitch “class” i.e., pitches

that are a multiple-of-octaves apart are considered equivalent. For HEAR 2021 we created two versions of this dataset: a 5 hour and 50 hour version. Unlike Carr et al. (2021), we treat this as a classification, not regression problem.

DCASE 2016 Task 2 A novel office sound event detection in synthesized scenes, adapted from DCASE 2016 Task 2 (Mesaros et al., 2018). Novel, insofar as our evaluation uses different splits. The original imbalanced splits did not work well our generic cross-validation.

Postprocessing: Predictions were postprocessed using 250 ms median filtering. At each validation step, a minimum event duration of 125 or 250 ms was chosen to maximize onset-only event-based F-measure (with 200 ms tolerance). Scores were computed using `sed_eval` (Mesaros et al., 2016).

Beehive States This is a binary classification task using audio recordings of two beehives (Nolasco et al., 2019). The beehives are in one of two states: a normal state, and one in which the queen bee is missing (“queen-less”). At 10 minutes long, this task has the longest audio clips in HEAR.

Beijing Opera Percussion This is a novel audio classification task developed using the Beijing Opera Percussion Instrument Dataset (Tian et al., 2014). The Beijing Opera uses six main percussion instruments that can be classified into four main categories: Bangu, Naobo, Daluo, and Xiaoluo.

CREMA-D CREMA-D is a dataset for emotion recognition (Cao et al., 2014). The original dataset contains audiovisual data of actors reciting sentences with one of six different emotions (anger, disgust, fear, happy, neutral and sad). For HEAR 2021, we only use the audio recordings (which differs from much but not all of the literature).

ESC-50 This is a multiclass classification task on environmental sounds. The ESC-50 dataset is a collection of 2000 environmental sounds organized into 50 classes (Piczak, 2015). Scores are averaged over 5 folds. (The folds are predefined in the original dataset.)

FSD50K FSD50K is a multilabel task (Fonseca et al., 2020). This dataset contains over 100 hours of human-labeled sound events from Freesound (<https://freesound.org/>). Each of the ≈ 51 k audio clips is labeled using one or more of 200 classes from the AudioSet Ontology, encompassing environmental sounds, speech, and music. Unlike the other datasets, for FSD50K scene embeddings we did not alter the audio clip length. Each clip is between 0.3 and 30 seconds long. We use the predefined train/val/eval split. Evaluation is done using mean average precision (mAP).

Gunshot Triangulation Gunshot triangulation is a novel resource multiclass classification task that utilizes a unique dataset: gunshots recorded in an open field using iPod Touch devices (Cooper and Shaw, 2020). This data consist of 22 shots from 7 different firearms, for a total of 88 audio clips, the smallest dataset in HEAR. Each shot is recorded using four different iPod Touches, located at different distances from the shooter. The goal of this task is to classify audio by the iPod Touch that recorded it, i.e., to identify the location of the microphone. The dataset was split into 7 different folds, where each firearm belonged to only one fold. Results are averaged over each fold.

GTZAN Genre The GTZAN Genre Collection (Tzanetakis and Cook, 2002) is a dataset of 1000 audio tracks (each 30 seconds in duration) that are categorized into ten genres (100 tracks per genre). The task is multiclass classification. As per the literature, scores are averaged over 10 folds. However, we don't used the corrected artist-conditional splits from (Sturm, 2013).

GTZAN Music Speech **GTZAN Music Speech** is a binary classification task, where the goal is to distinguish between music and speech. The dataset consists of 120 tracks (each 30 seconds in duration) and each class (music/speech) has 60 examples.

LibriCount LibriCount is a multiclass speaker count identification task (Stöter et al., 2018b). The dataset contains audio of a simulated cocktail party environment with between 0 to 10 speakers. The goal of this task is to classify how many speakers are present in each of the recordings. Following Stöter et al. (2018a), we treat this as a classification, not regression, problem.

MAESTRO 5h This is a novel music transcription task adapted from MAESTRO. For HEAR 2021, we created a subsampled version that includes 5 hours of training and validation audio, in 120 second clips. To evaluate submissions, a shallow transcription model was trained on timestamp-based embeddings provided by the participant models.

We use note onset FMS and note onset with offset FMS for evaluation, as per the original MAESTRO paper (Hawthorne et al., 2019) and the preceding Onsets and Frames paper (Hawthorne et al., 2018).

Note onset measures the ability of the model to estimate note onsets with 50 ms tolerance and ignores offsets. Note onset w/ offset includes onsets as well as requires note duration within 20% of ground truth or within 50 ms, whichever is greater.

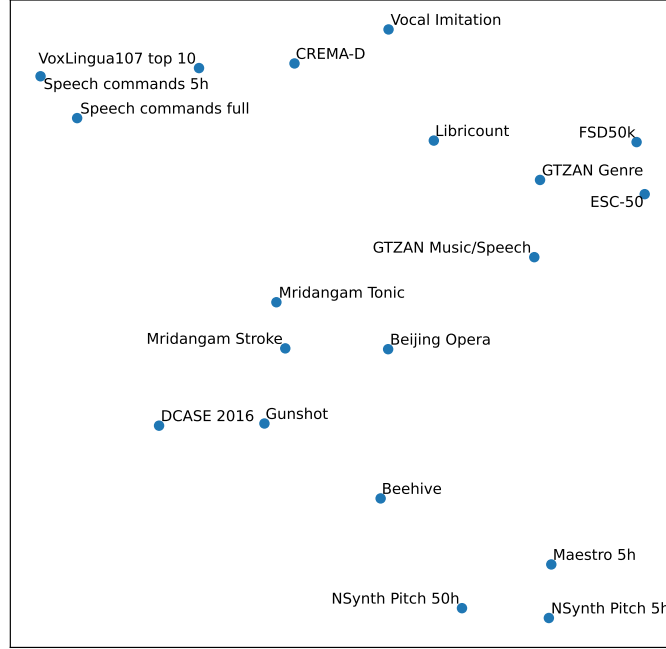
Mridingham Stroke and Mridingham Tonic We used the Mridangam Stroke Dataset (Anantapadmanabhan et al., 2013) for two novel multiclass classification tasks: Stroke classification and Tonic classification. The Mridingam is a pitched percussion instrument used in carnatic music, which is a sub-genre of Indian classical music. This dataset comprises 10 different strokes played on Mridingams with 6 different tonics.

Vocal Imitations Vocal Imitations (Kim et al., 2018a) is a novel multiclass classification task, where the goal is to match a vocal imitation of a sound with the sound that is being imitated. The dataset contains 5601 vocal imitations of 302 reference sounds, organized by AudioSet ontology. Given a vocal sound, the classification task is to retrieve the original audio it is imitating.

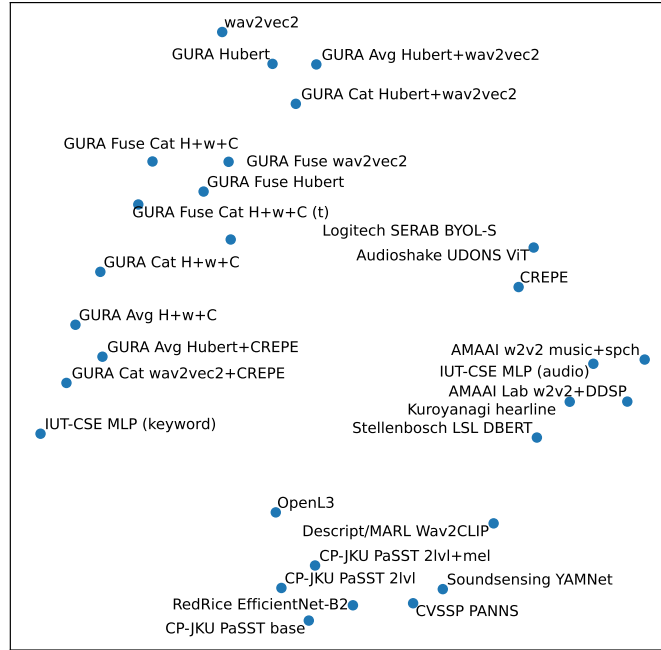
VoxLingua107 Top 10 This is a novel multiclass classification task derived from the VoxLingua107 dataset (Valk and Alumäe, 2021). The goal of the task is to identify the spoken language in an audio file. For HEAR 2021 we selected the top 10 most frequent languages from the development set, which resulted in just over 5 hours of audio over 972 audio clips.

Table 3: Properties of baseline and submitted models, including: whether the model processes raw audio (1-D) or spectrograms (2D); on what kind of data the model is pretrained; the number of million parameters; the size of the output embedding for scene and timestamp tasks; and the number of minutes the model spends embedding Speech Commands V2. We caution that embedding time is not the entire picture, if participants did not do simple speed optimizations. For example, the CREPE wrapper (also used by GURA) is known not to exploit GPU batch parallelism.

Model	Input		Pretraining data			# M params	Embed dim		Time min
	1D	2D	speech	broad	music		scene	time	
OpenL3		✓		✓		4.7	512	512	94.9
wav2vec2	✓		✓			315.4	1024	1024	8.9
CREPE	✓				✓	22.2	2048	2048	38.3
AMAAI Lab wav2vec2+DDSP	✓		✓		✓	98.8	871	871	43.6
AMAAI wav2vec2 music+speech	✓		✓		✓	300.0	768	768	5.0
Audioshake UDONS ViT		✓	✓			11.1	768	768	3.5
CP-JKU PaSST 2lvl		✓		✓		86.2	1295	2590	14.5
CP-JKU PaSST 2lvl+mel		✓		✓		86.2	1295	3358	5.8
CP-JKU PaSST base		✓		✓		86.2	1295	1295	5.8
CVSSP PANNS		✓		✓		80.8	2048	2048	3.9
Descript/MARL Wav2CLIP		✓		✓		11.7	512	512	3.1
GURA Avg H+w+C	✓		✓		✓	1339.0	1024	1024	40.0
GURA Avg Hubert+Crepe	✓		✓		✓	1022.0	1024	1024	33.9
GURA Avg Hubert+wav2vec2	✓		✓			634.0	1024	1024	14.6
GURA Cat H+w+C	✓		✓		✓	1339.0	3072	3072	40.1
GURA Cat Hubert+wav2vec2	✓		✓			634.0	2048	2048	14.4
GURA Cat wav2vec2+crepe	✓		✓		✓	339.0	2048	2048	24.7
GURA Fuse Cat H+w+C	✓		✓		✓	1339.0	3072	3072	40.1
GURA Fuse Cat H+w+C (time)	✓	✓			✓	1339.0	15360	3072	34.6
GURA Fuse Hubert	✓		✓			1000.0	1280	1280	18.1
GURA Fuse wav2vec2	✓		✓			317.0	1024	1024	8.8
GURA Hubert	✓		✓			1000.0	1280	1280	17.9
IUT-CSE MLP (audio)		✓	✓	✓	✓	0.2	1584	8	2.9
IUT-CSE MLP (keyword)		✓	✓			0.4	1024	64	3.0
Logitech AI SERAB BYOL-S		✓	✓			5.3	2048	2048	4.8
RedRice EfficientNet-B2		✓		✓		7.7	1408	1408	3.4
Soundsensing YAMNet		✓		✓		3.8	1024	1024	15.7
Stellenbosch LSL DBERT	✓		✓			316.8	2048	2048	6.5



(a) Tasks



(b) Models

Figure 2: t-SNE visualizations of tasks and models, based upon normalized scores. Missing normalized scores were imputed using sklearn’s multivariate IterativeImputer.

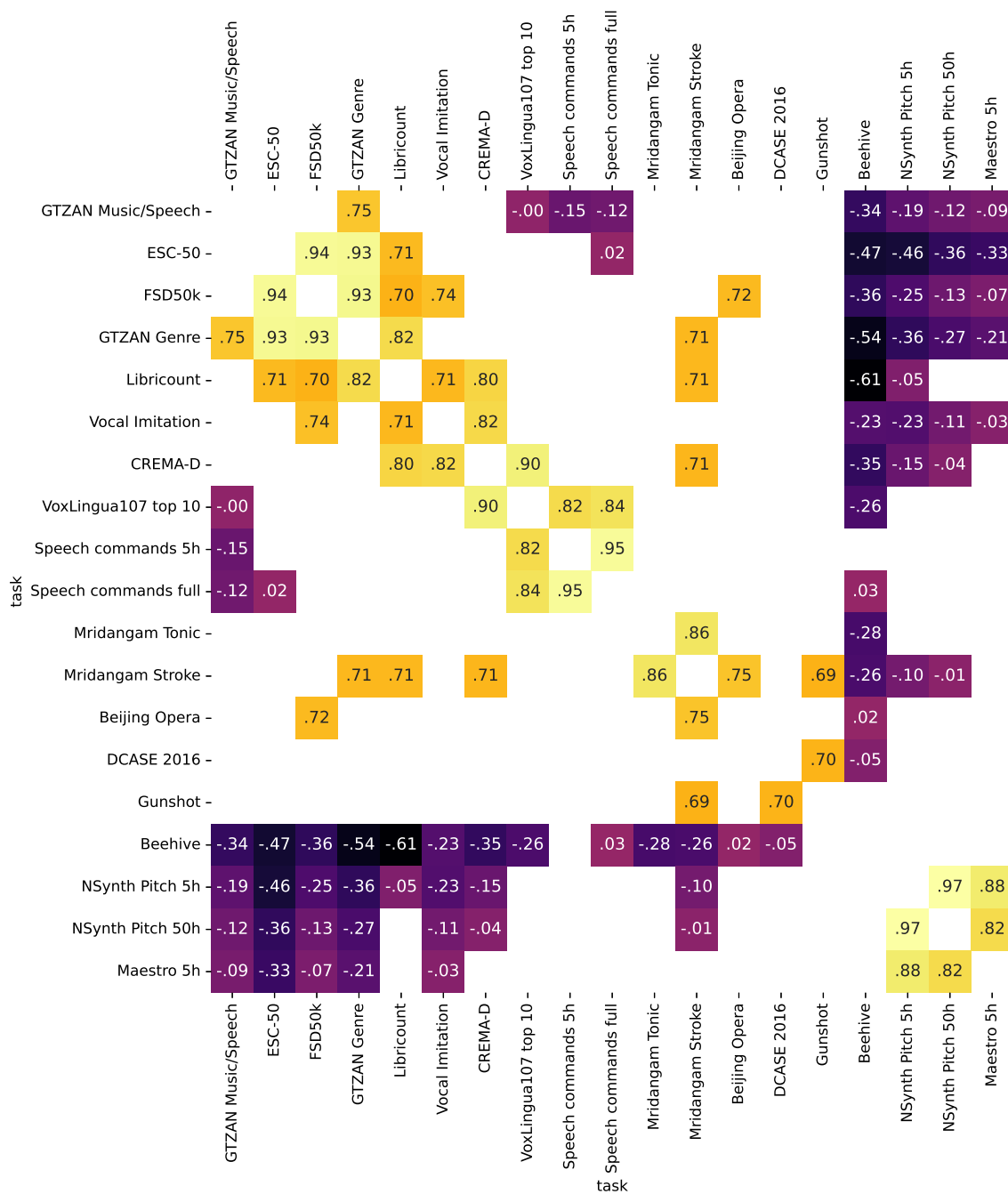


Figure 3: Task versus task correlation scores, based upon normalized scores. Only the highest and lowest correlations are displayed. Cells are sorted to minimize the traveling salesperson distance, mapping correlations $[-1, +1]$ to distances $[+2, 0]$.

HEAR 2021: HOLISTIC EVALUATION OF AUDIO REPRESENTATIONS

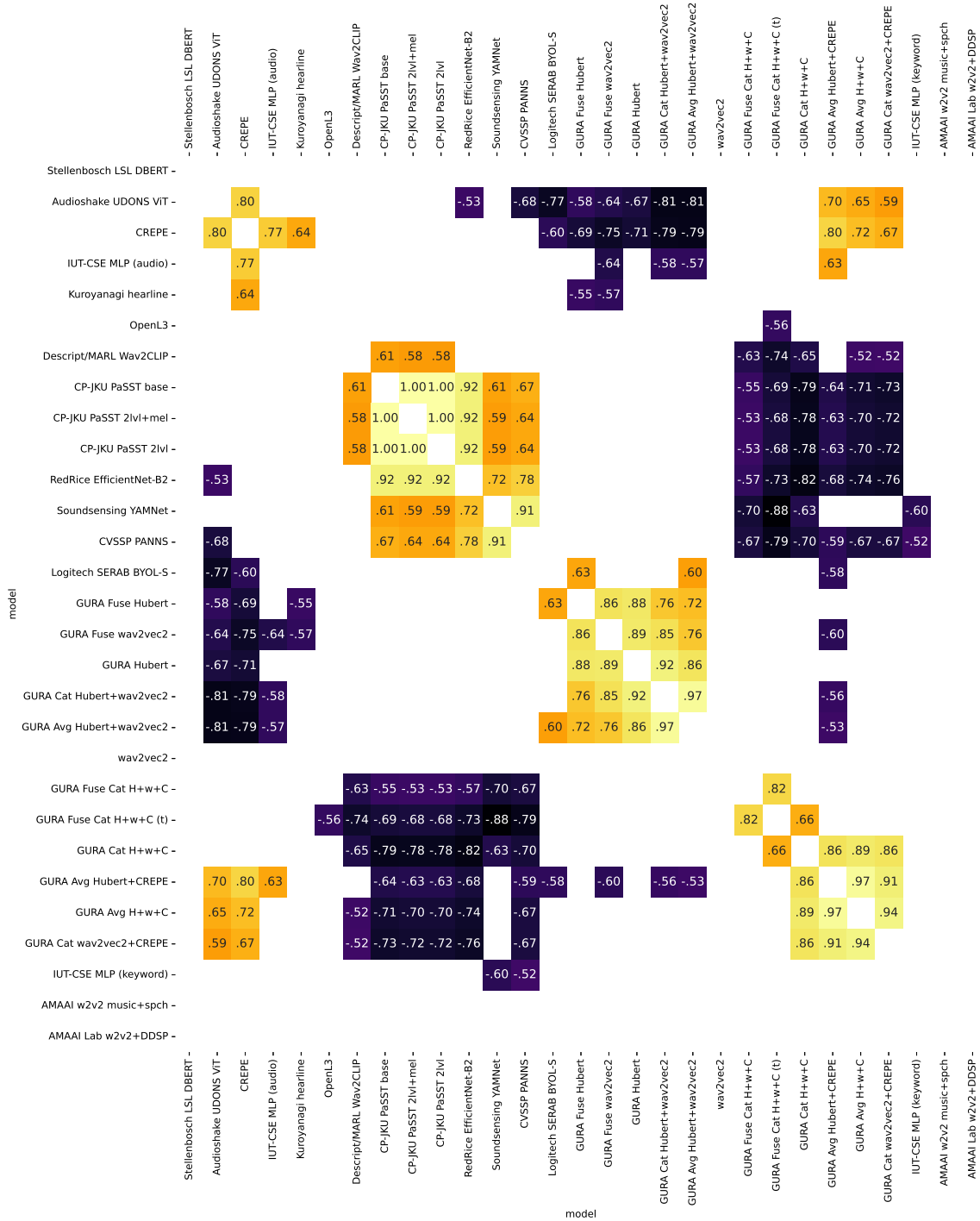


Figure 4: Model versus model correlation scores, based upon normalized scores. Only the highest and lowest correlations are displayed. Cells are sorted to minimize the traveling salesperson distance, mapping correlations [-1, +1] to distances [+2, 0].

Appendix B. Downstream training details

For each task, using a given model’s frozen embeddings as input features, we train a downstream MLP classifier. For scene-based multiclass tasks, the final layer is a softmax with cross-entropy loss. For scene-based multilabel tasks and multilabel frame reductions of timestamp tasks, the final layer is a sigmoid with cross-entropy loss.

We monitor the score (not loss) on the validation set. For timestamp tasks, computing the validation score involves a full CPU-based `sed_eval` (Mesaros et al., 2016) run with median filter of 250ms and minimum event duration 125 ms and 250 ms. (Both event durations are tried at each validation step and the best hyperparameter is retained for that validation step.) We train for a maximum of 500 epochs, checking the validation score every 3 epochs, early stopping if no improvement is seen after 20 validation steps. For DCASE 2015 Task 2, we check the validation score every 10 epochs.

The validation score is used for early-stopping, as well as for model selection. The same RNG seed is used for every model-task downstream training, ensuring that grid points and weight initialization is identical. Model selection is performed over 8 deterministic random grid points out of 16 possible grid points. Hyperparameters are shown in Table 4. This grid was chosen after using a much larger hyperparameter grid with the three baseline models on the open tasks. In these preliminary hyperparameter grid pruning experiments, the grid was progressively refined by discarding hyperparameter choices that were not predictive of relatively high model performance, similarly to how Kelz et al. (2016) use tree ensemble learning to prune their hyperparameter grid.

Table 4: Hyperparameters used for training.

Hidden layers	[1, 2]
Hidden dimensions	1024
Dropout	0.1
Learning rate	[3.2e-3, 1e-3, 3.2e-4, 1e-4]
Batch size	1024
Hidden norm	Batch Norm
Initialization	[Xavier Uniform, Xavier Normal] (Glorot and Bengio, 2010)
Optimizer	Adam