On the Transferability of Large-Scale Self-Supervision to Few-Shot Audio Classification

ON THE TRANSFERABILITY OF LARGE-SCALE SELF-SUPERVISION TO FEW-SHOT AUDIO CLASSIFICATION

Calum Heggan[1]*, Sam Budgett*[2]*, Tim Hospedales* [1]*, Mehrdad Yaghoobi*[1]

[1] University of Edinburgh, Scotland, [2] Thales UK RTI

ABSTRACT

In recent years, self-supervised learning has excelled for its capacity to learn robust feature representations from unlabelled data. Networks pretrained through self-supervision serve as effective feature extractors for downstream tasks, including Few-Shot Learning. While the evaluation of unsupervised approaches for few-shot learning is well-established in imagery, it is notably absent in acoustics. This study addresses this gap by assessing large-scale self-supervised models' performance in few-shot audio classification. Additionally, we explore the relationship between a model's few-shot learning capability and other downstream task benchmarks. Our findings reveal state-of-the-art performance in some few-shot problems such as SpeechCommandsv2, as well as strong correlations between speech-based few-shot problems and various downstream audio tasks.

Index Terms— Self-Supervision, Few-Shot Learning

1. INTRODUCTION

Both few-shot learning and self-supervised learning have become increasingly popular in response to the lack of large labelled datasets in many domains and practical applications [1, 2]. Models pre-trained using self-supervision, the act of generating and solving self-generated tasks, have demonstrated strong success on few-shot learning tasks. Despite significant strides in other areas of benchmarking these approaches for audio problems questions persist regarding few-shot capabilities. Specifically, the effectiveness of self-supervision approaches for downstream few-shot adaptation is unclear due to the diverse range of methods and architectures employed in published models. This is further complicated by the substantial computational costs required to train large-scale models from scratch. The alignment of model rankings based on few-shot performance with those for other tasks is a critical consideration, influencing whether progressions from other audio-related tasks can be leveraged for few-shot learning. If misaligned, it may signal the need to include few-shot tasks in future holistic audio self-supervision benchmarks.

This work is supported by the Engineering and Physical Sciences Research Council of the UK (EPSRC) Grant number EP/S000631/1 and the UK MOD University Defence Research Collaboration (UDRC) in Signal Processing, EPSRC iCASE account EP/V519674/1 and Thales UK Ltd.

To address these issues, we conduct an extensive evaluation of state-of-the-art (SOTA) pre-trained self-supervised models for downstream few-shot audio classification. Our study includes 13 pre-trained models evaluated across 10 diverse few-shot datasets, spanning environmental, animal, and speech sounds. Key contributions include, a) identifying the most effective approach for few-shot audio classification, b) understanding differences in algorithm ranking between few-shot and other benchmarks, and c) exploring relationships between few-shot and other tasks.

2. RELATED WORK

2.1. Few-Shot Learning

Few-shot learning aims to learn a task from limited labelled examples. Various fields address this challenge, with meta-learning being a prominent approach. Meta-learning involves generating and solving similarly structured few-shot tasks by leveraging a labelled training dataset. Two major groups emerge: I) Gradient-Based Meta-Learning (GBML), where models adapt rapidly to new tasks, and II) Metric-based approaches, which learn an embedding network [1]. Classical pre-training with labelled base classes [1] and, more recently, self-supervision, where models are trained on an unlabelled dataset using pre-text tasks, have also shown success [2]. While algorithms for few-shot learning are prevalent in the image domain, their exploration in the audio domain is limited, with only a few being reproducible. We heavily rely on the MetaAudio benchmark [3] and its extension from MT-SLVR [4] for our few-shot evaluation. This benchmark provides clear testing settings and diverse meta-learning paradigms for the evaluation tasks.

2.2. Self-Supervision

Self-supervised representation learning is a large and rich topic, both within acoustics and other modalities. As such we refer readers to more comprehensive surveys of available approaches and their unique attributes [5]. Here, we instead focus on relevant families of algorithms. A popular family of approaches is prediction-based self-supervision, where a model is trained to predict the context of unseen sections

of data [6–11], or to contrast the unseen target frame with randomly sampled ones [12, 13]. Multi-task approaches combining multiple objectives have also been explored [4, 14]. Discrete target methods, like clustering in HuBERT and DistilHuBERT [15, 16], also contribute to this rich landscape.

2.3. Benchmarks & Evaluations

With increasing number of proposed approaches, the value of rigorous empirical evaluation has grown [2]. Within few-shot classification, evaluation has largely been around imagery, however more recently has spanned into other domains. Of particular relevance to this work are the MetaAudio [3] benchmark and its follow-up MT-SLVR [4], which to date contain the most diverse set of downstream few-shot audio classification tasks. Within self-supervision, there has been a rapid growth of benchmarking efforts. Our work is most closely related to those focused on acoustics, the most ubiquitous of which is the Speech Processing Universal PERformance Benchmark (SUPERB) [17]. In total, SUPERB comprises 11 unique tasks but despite being diverse within the speech domain, lacks evaluation of few-shot capabilities, the ability to perform which is vital in many practical problems.

Table 1: Summary of pre-trained models. Abbreviations following [17]: (VQ) vector quantization, (F) future, (M) masked, (G) generation, (C) contrastive, (P) token prediction/classification, (UM) utterance mixing and (GREP) gated relative position bias. Param count includes pre-training and inference.

Approach	N^o Params (M)	Objective(s)	Network	Input
WavLM Base [11]	94.38	M-P + VQ + GREP + UM	7-Conv 12-Trans	Raw
HuBERT Base [15]	94.68	M-P + VQ	7-Conv 12-Trans	Raw
wav2vec 2.0 Base [13]	95.04	M-C + VQ	7-Conv 12-Trans	Raw
DistilHuBERT [16]	21.32	Layer Distillation	7-Conv 2-Trans	Raw
DeCoAR 2.0 [8]	89.84	M-G + VQ	12-Trans	Raw
wav2vec [12]	32.54	F-C	19-Conv	Raw
vq-wav2vec [18]	34.15	F-C + VQ	20-Conv	Raw
APC [6]	4.11	F-G	3-GRU	Raw
VQ-APC [19]	4.63	F-G + VQ	3-GRU	Raw
NPC [20]	19.38	M-G + VQ	4-Conv, 4-Masked Conv	Raw
TERA [10]	21.33	Time/Freq M-G	3-Trans	Spec
PASE+ [14]	7.83	Multi-Task	SincNet, 7-Conv, 1-QRNN	Raw
MockingJay [9]	21.33	Time M-G	12-Trans	Spec

Table 2: Summary of few-shot datasets: Speech (top), environmental (middle), and animal (bottom) sounds.

Name	Setting	N^o Classes	N^o Samples	Length (s)
VoxCeleb1	Speaker	1,251	153,516	3-180
SpeechCommandsV2	Keyword	35	105,829	1
Crema-D	Emotion Recognition	6	7,442	1s - 5
Speech Accent Archive	Accent	122	2,060	17s - 110
Common Voice v12 Delta	Language	88	256,243	530
ESC-50	Environmental	50	2,000	5
NSynth	Instrumentation	1,006	305,978	4
FDSKaggle18	Mixed	41	11,073	0.3-30
Watkins Marine Mammal	Marine Mammals	32	1,698	0.1-150
BirdCLEF 2020 (Pruned)	Bird Song	715	63,364	3-180

3. SELF-SUPERVISION FOR FEW-SHOT LEARNING

Self-supervised pre-training is a strong candidate for few-shot cases where a large quantity of unlabelled data is available. In such cases, we are able to learn general purpose representations which can then be used directly or updated with few samples. We study a pipeline in which trained models are frozen and used as feature extractors. Then for each task we train a lightweight linear classifier. We evaluate the included approaches on few-shot audio classification tasks. Such few-shot tasks are individual learning problems, each containing a small training set (support set $\mathcal{S}$) and testing set (query set $\mathcal{Q}$). Few-shot tasks are commonly expressed as N-Way K-Shot tasks, where N is the number of classes being discriminated between, while K is the number of labelled examples per class available. Formally, N-Way K-Shot tasks take the form:

$$\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{\mathcal{M}}, y_{\mathcal{M}})\} \quad (1)$$

$$\mathcal{Q} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{\mathcal{L}}, y_{\mathcal{L}})\} \quad (2)$$

where each (x, y) pair consists of an input $\mathbf{x} \in \mathbb{R}^D$ and a class label $\mathbf{y} \in \{1, \ldots, N\}$, and where $\mathcal{M}$ and $\mathcal{L}$ are the total number of support and query examples respectively.

4. SETUP

4.1. Models & Pre-Training

We utilise models which have been submitted and evaluated on the SUPERB benchmark, and that are available through the s3prl toolkit [9, 10]. To keep pipelines as fair as possible, we only evaluate models which have been pre-trained using LibriSpeech 960 (LS960). In total, we consider 13 models spanning a variety of training objectives, as detailed in Table 1. We refer the reader to the original works for implementation details.

4.2. Few-Shot Evaluation & SUPERB

We conduct evaluation of few-shot audio classification across ten datasets, encompassing speech, environmental, and animal sounds (details in Table 2). We incorporate few-shot tasks beyond the speech pre-trained domain for two reasons: I) to assess the extent to which tasks based on animal sounds, somewhat similar to human speech, can be evaluated using speech features, and II) to gauge the out-of-domain transfer performance of speech models to other audio forms. For datasets designed for meta-learning, with class-wise splits, we exclusively use the provided test classes. This constraint ensures a fair comparison with previous works such as MetaAudio and MT-SLVR [3,4]. During evaluation, pre-trained models are frozen and employed as feature extractors, with new linear classifiers trained per task. All included models output 2D features, comprising both a traditional feature dimensions

Table 3: Average percentage accuracy for few-shot audio classification. Each result is the mean and 95% confidence interval of 10,000 random 5-way 1-shot tasks. We also include current SOTA results. Results style: **Best**, **Second Best**.

Method	Speech					Enviromental			Animal			
	SCv2	SAA	CommonVoice	VoxCeleb	Crema-D	NSynth	ESC-50	Kaggle18	Watkins	BirdClef	Avg	Avg Rank
WavLM Base	$52.27^{\pm0.43}$	$\mathbf{26.92}^{\pm0.33}$	$\mathbf{31.72}^{\pm0.38}$	$27.68^{\pm0.36}$	$27.86^{\pm0.38}$	$56.69^{\pm0.42}$	$48.29^{\pm0.40}$	$35.85^{\pm0.40}$	$43.29^{\pm0.42}$	$28.26^{\pm0.37}$	$37.88^{\pm0.07}$	7.1
HuBERT Base	$\underline{\mathbf{57.10}}^{\pm0.43}$	$26.30^{\pm0.34}$	$31.34^{\pm0.38}$	$28.36^{\pm0.37}$	$28.17^{\pm0.38}$	$61.45^{\pm0.41}$	$\mathbf{55.41}^{\pm0.41}$	$37.71^{\pm0.41}$	$44.25^{\pm0.43}$	$30.30^{\pm0.38}$	$\mathbf{40.04}^{\pm0.07}$	4.7
wav2vec 2.0 Base	$34.41^{\pm0.39}$	$25.67^{\pm0.34}$	$30.02^{\pm0.37}$	$27.30^{\pm0.36}$	$\underline{\mathbf{29.91}}^{\pm0.37}$	$50.09^{\pm0.42}$	$46.90^{\pm0.42}$	$33.20^{\pm0.39}$	$41.34^{\pm0.42}$	$28.04^{\pm0.37}$	$34.69^{\pm0.06}$	9.5
DistilHuBERT	$\mathbf{55.20}^{\pm0.43}$	$25.98^{\pm0.34}$	$31.63^{\pm0.38}$	$28.27^{\pm0.37}$	$28.80^{\pm0.38}$	$60.17^{\pm0.41}$	$55.16^{\pm0.41}$	$36.60^{\pm0.41}$	$45.47^{\pm0.42}$	$29.70^{\pm0.37}$	$\mathbf{39.70}^{\pm0.07}$	4.9
DeCoAR 2.0	$37.05^{\pm0.40}$	$24.19^{\pm0.34}$	$30.32^{\pm0.38}$	$30.62^{\pm0.38}$	$26.72^{\pm0.37}$	$66.95^{\pm0.39}$	$49.14^{\pm0.41}$	$34.24^{\pm0.38}$	$43.23^{\pm0.42}$	$\underline{\mathbf{30.87}}^{\pm0.37}$	$37.33^{\pm0.06}$	7.4
wav2vec	$41.08^{\pm0.41}$	$22.15^{\pm0.33}$	$30.88^{\pm0.38}$	$28.25^{\pm0.37}$	$27.83^{\pm0.38}$	$49.83^{\pm0.42}$	$51.21^{\pm0.41}$	$34.58^{\pm0.39}$	$42.32^{\pm0.41}$	$28.00^{\pm0.37}$	$35.61^{\pm0.05}$	9.4
vq-wav2vec	$41.06^{\pm0.41}$	$22.04^{\pm0.31}$	$27.88^{\pm0.37}$	$26.60^{\pm0.36}$	$28.86^{\pm0.37}$	$50.40^{\pm0.41}$	$48.34^{\pm0.39}$	$32.36^{\pm0.38}$	$37.76^{\pm0.42}$	$27.78^{\pm0.37}$	$34.31^{\pm0.07}$	11.2
APC	$42.01^{\pm0.41}$	$22.42^{\pm0.34}$	$31.01^{\pm0.38}$	$\mathbf{32.47}^{\pm0.39}$	$29.45^{\pm0.38}$	$64.38^{\pm0.40}$	$52.77^{\pm0.42}$	$36.12^{\pm0.40}$	$46.28^{\pm0.43}$	$\mathbf{30.50}^{\pm0.37}$	$38.74^{\pm0.06}$	**4.5**
VQ-APC	$38.95^{\pm0.40}$	$24.72^{\pm0.34}$	$29.46^{\pm0.37}$	$29.83^{\pm0.38}$	$28.56^{\pm0.37}$	$63.37^{\pm0.40}$	$49.89^{\pm0.41}$	$34.50^{\pm0.38}$	$43.23^{\pm0.41}$	$27.05^{\pm0.35}$	$36.96^{\pm0.07}$	8.4
NPC	$31.18^{\pm0.37}$	$21.65^{\pm0.31}$	$27.55^{\pm0.35}$	$28.40^{\pm0.35}$	$27.49^{\pm0.35}$	$59.44^{\pm0.40}$	$46.26^{\pm0.40}$	$33.05^{\pm0.37}$	$43.12^{\pm0.41}$	$27.90^{\pm0.35}$	$34.60^{\pm0.05}$	11.6
TERA	$32.56^{\pm0.43}$	$24.30^{\pm0.33}$	$30.21^{\pm0.38}$	$30.47^{\pm0.36}$	$29.24^{\pm0.38}$	$69.45^{\pm0.42}$	$52.47^{\pm0.40}$	$35.11^{\pm0.40}$	$\mathbf{50.68}^{\pm0.42}$	$29.85^{\pm0.37}$	$38.43^{\pm0.05}$	5.6
PASE+	$30.00^{\pm0.37}$	$24.70^{\pm0.34}$	$30.76^{\pm0.36}$	$26.86^{\pm0.34}$	$26.50^{\pm0.36}$	$37.36^{\pm0.39}$	$52.20^{\pm0.40}$	$\underline{\mathbf{39.20}}^{\pm0.40}$	$45.30^{\pm0.42}$	$28.66^{\pm0.34}$	$34.15^{\pm0.07}$	8.7
Mockingjay	$29.79^{\pm0.38}$	$23.58^{\pm0.33}$	$29.65^{\pm0.37}$	$28.87^{\pm0.37}$	$28.61^{\pm0.38}$	$\mathbf{70.71}^{\pm0.38}$	$46.28^{\pm0.42}$	$33.52^{\pm0.39}$	$47.23^{\pm0.43}$	$27.61^{\pm0.36}$	$36.58^{\pm0.05}$	8.9
MT-SLVR (SOTA) [4]	$23.65^{\pm0.34}$	$\underline{\mathbf{28.92}}^{\pm0.37}$	$\underline{\mathbf{35.22}}^{\pm0.40}$	$\underline{\mathbf{33.58}}^{\pm0.39}$	$\mathbf{29.61}^{\pm0.38}$	$\underline{\mathbf{71.81}}^{\pm0.39}$	$\underline{\mathbf{69.53}}^{\pm0.39}$	$\mathbf{38.36}^{\pm0.40}$	$\underline{\mathbf{59.49}}^{\pm0.42}$	$29.49^{\pm0.38}$	$\underline{\mathbf{41.97}}^{\pm0.02}$	$\underline{\mathbf{3.0}}$

and a time dimension. Since this 2D representation isn't suitable for linear classifiers, we collapse one of the dimensions, with the method of doing so treated as a hyperparameter. We experimented with min, max and average pooling, and found that averaging the time dimension resulted in the highest validation performance. We also use the SUPERB benchmark, which comprises 11 tasks spanning 4 aspects of speech: content, speaker, semantics, and paralinguistics [17].

4.3. Correlation

To analyse the relationship between downstream few-shot and upstream SUPERB benchmarks, we use Pearson Rank correlation. To account for relative importance of absolute changes, we utilise the logit transformation of all metrics that have a range between 0 and 1 [2]. This includes all accuracies and SUPERB measures except for model ranks score. Where possible, bootstrap resampled correlation errors are given.

4.4. Limitations

Due to availability of pre-trained models and compute resources, our work has limitations. Many SOTA audio methods often use unique architectures. This, together with the immense cost of training from scratch, means that our comparison cannot account for differences in backbone. We are also limited to linear readout evaluation due to fine-tuning cost.

5. RESULTS

5.1. Few-Shot Performance

The few-shot results in Table 3 reveal a variety of noteworthy insights. Firstly, we note significantly improved performance in few-shot keyword spotting with SpeechCommandsV2 (SCv2) in almost all included models, with the highest performance being achieved by HuBERT Base. Intriguingly, pre-training with LS960, while effective for few-shot keywords, appears to yield suboptimal results for other included speech tasks compared to pre-training on environmental sounds [4]. For Kaggle18, NSynth and BirdClef, we achieve competitive or SOTA performance, an interesting result given that these sets are not based in speech. This result suggests that there is a possible overlap between required representation space for instrumentation/birdsong and speech. Although not competitive with SOTA approaches, performance on underwater mammal classification does overlap with many results from the joint training meta-learning condition in MetaAudio [3]. Averaging over all sets, we observe that HuBERT base, followed closely by DistilHuBERT, performs best. Using average rank, we observe that APC performs consistently well, although suffering from poorer relative performance on a few key sets such as SCv2.

5.2. Relationship to SUPERB

We analyse the relationship between SUPERB and few-shot tasks by considering task-wise correlation Figures 1 and 2. We observe that speech sets, with the exclusion of VoxCeleb and Crema-D, exhibit significantly higher correlation across all tasks compared to environmental and animal sounds. Surprisingly, despite being rooted in speech, VoxCeleb and Crema-D display weak correlations with their traditional evaluation counterparts. This particularly interesting given that the same VoxCeleb dataset is used in SUPERBs' speaker identification task. In addition to its relative weakness, we also observe that VoxCeleb shows an unexpected negative relationship with considered tasks, e.g. as Automatic Speech Recognition performance gets worse, few-shot VoxCeleb performance improves. Interestingly, underwater mammals (Watkins) mirrors this behaviour. Environmental sets exhibit consistently low correlation with SUPERB tasks, while animal sounds achieve a moderately stronger relationship. For few-shot tasks that correlate well with any SUPERB task, correlation is strong across all tasks. The strongest single correlation is observed between Query-by-Example (QBE) and SCv2. Looking at the SUPERB benchmark more holistically, we can model the relationship between our few-shot tasks and SUPERB's model score, Figure 1. Even in subdomain

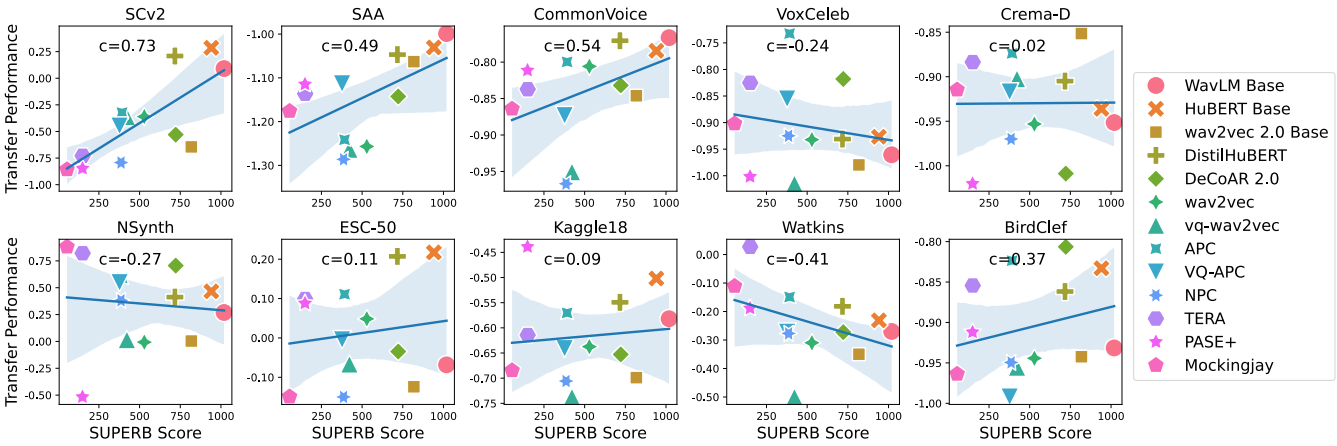

Fig. 1: SUPERB model score vs average few-shot transfer performance for all considered datasets. (TOP) row contains speech datasets, (BOTTOM) row contains environmental/animal sets. Regression gradients and shaded regions describe correlation strength and 95% confidence intervals respectively. Spearman Rank correlation coefficients (c) are shown top left of each plot.

Fig. 2: Spearman rank correlations between Few-Shot (rows) and SUPERB (cols) tasks. Few-shot tasks are split into speech (top), environment (mid) and animal (bottom) sounds. SUPERB is split into context, speaker, semantics and paralinguistics (left to right).

few-shot tasks such as few-shot speech, evidence suggests that performance is only aligned to SUPERB's scores in a few cases. For included speech problems, we note that performance variation over models is also fairly low compared to variation in SUPERB tasks, a possible extrapolation of which could suggests high specialisation in many of the included tasks gives only marginal gains on few-shot tasks. We also note that one reason for potential low correlation in many tasks is how features from these models are used downstream in SUPERB. Although the model itself is never fine-tuned, the linear model used on top of the frozen extractors typically has access to many more samples in SUPERB tasks than would be found in few-shot tasks. How different the features of a non data-constrained linear model for a given SUPERB task are from the base model features may have a significant impact. For robust future performance and low-data aligned SOTA models, firstly we propose that speech based few-shot tasks should be included in SUPERB style benchmarks for overall model scoring. In addition to adding speech based tasks, secondly we suggest that adding some additional related tasks such as animal sounds and instrumentation could be beneficial, due to their seemingly similar qualities.

6. CONCLUSION

In this study, we investigated the efficacy of large-scale self-supervised models in the realm of few-shot audio and speech classification, concurrently exploring their relationship with the widely-used self-supervised audio benchmark SUPERB. Our comprehensive evaluation of 13 models across 10 diverse few-shot audio datasets revealed notable insights, including the establishment of a new state-of-the-art for SCv2 and the limited impact of speech pre-training on few-shot speech tasks. Three out of five speech-based few-shot tasks demonstrated high correlations with SUPERB tasks, indicating potential domain connections. Animal sounds also exhibited some relation to SUPERB. Conversely, negative relationships were observed between many few-shot and SUPERB tasks, suggesting conflicts between few-shot performance and existing benchmarks. Our study implies that SUPERB benchmark performance improvements may not generalise to low-resource settings. Consequently, we propose the inclusion of few-shot tasks, especially speech-related ones, in future speech self-supervision benchmarks for a more comprehensive evaluation of model capabilities in diverse scenarios.

7. REFERENCES

[1] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[2] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales, "How well do self-supervised models transfer?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[3] Calum Heggan, Sam Budgett, Timothy M. Hospedales, and Mehrdad Yaghoobi, "Metaaudio: A few-shot audio classification benchmark," in *ICANN*, 2022.

[4] Calum Heggan, Tim Hospedales, Sam Budgett, and Mehrdad Yaghoobi, "MT-SLVR: Multi-Task Self-Supervised Learning for Transformation In(Variant) Representations," in *Proc. INTERSPEECH 2023*, 2023.

[5] Abdel rahman Mohamed, Hung yi Lee, Lasse Borgholt, Jakob Drachmann Havtorn, Joakim Edin, C. Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[6] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass, "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Proc. Interspeech 2019*, 2019.

[7] Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee, "Audio albert: A lite bert for self-supervised learning of audio representation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[8] Shaoshi Ling and Yuzong Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," 2020.

[9] Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6419–6423.

[10] Andy T. Liu, Shang-Wen Li, and Hung-yi Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021.

[11] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, 2022.

[12] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Interspeech 2019*, 2019.

[13] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, 2020.

[14] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[15] 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*, vol. 29, 2021.

[16] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[17] 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 *Proc. Interspeech 2021*, 2021.

[18] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2019.

[19] Yu-An Chung, Hao Tang, and James Glass, "Vector-quantized autoregressive predictive coding," 2020.

[20] Alexander H Liu, Yu-An Chung, and James Glass, "Non-autoregressive predictive coding for learning speech representations from local dependencies," 2021.