LightHuBERT: Lightweight and Configurable Speech Representation Learning with Once-for-All Hidden-Unit BERT

Rui Wang[1,†], Qibing Bai[2,†], Junyi Ao[3], Long Zhou[4], Zhixiang Xiong[1],
Zhihua Wei[1], Yu Zhang[2,5], Tom Ko[6], Haizhou Li[3]

[1]Department of Computer Science and Technology, Tongji University
[2]Department of Computer Science and Engineering, Southern University of Science and Technology
[3]School of Data Science, The Chinese University of Hong Kong (Shenzhen)
[4]Microsoft [5]Peng Cheng Laboratory [6]ByteDance AI Lab

Abstract

Self-supervised speech representation learning has shown promising results in various speech processing tasks. However, the pre-trained models, e.g., HuBERT, are storage-intensive Transformers, limiting their scope of applications under low-resource settings. To this end, we propose LightHuBERT, a once-for-all Transformer compression framework, to find the desired architectures automatically by pruning structured parameters. More precisely, we create a Transformer-based supernet that is nested with thousands of weight-sharing sub-nets and design a two-stage distillation strategy to leverage the contextualized latent representations from HuBERT. Experiments on automatic speech recognition (ASR) and the SU-PERB benchmark show the proposed LightHuBERT enables over 10^9 architectures concerning the embedding dimension, attention dimension, head number, feed-forward network ratio, and network depth. LightHuBERT outperforms the original HuBERT on ASR and five SUPERB tasks with the Hu-BERT size, achieves comparable performance to the teacher model in most tasks with a reduction of 29% parameters, and obtains a $3.5\times$ compression ratio in three SUPERB tasks, e.g., automatic speaker verification, keyword spotting, and intent classification, with a slight accuracy loss. The code and pre-trained models are available at `https://github.com/mechanicalsea/lighthubert`.

Index Terms: speech pre-training, model compression, knowledge distillation, neural architecture search, Transformer

1. Introduction

Self-supervised speech representation learning has shown that pre-trained models benefit from abundantly available unannotated data and produce promising results on various speech processing tasks [1, 2, 3, 4, 5, 6, 7]. Most speech applications require models to interact with humans or machines, therefore demanding real-time performance for a better user experience. However, many real-world devices, e.g., smartwatches, mobile phones, tablets, audio-visual robots, and industrial PCs, are highly constrained by memory and battery. This prevents the pre-trained speech models from actual deployments.

Building a compression framework for pre-trained models to meet various resource constraints can be essential for the development of speech pre-training. It can utilize pre-trained models trained on large-scale unlabeled data to boost the performance of compressed networks in downstream tasks and lower entry barriers for pre-training related works. However, two challenges remain when compressing pre-trained models without compromising their effectiveness: (1) a lightweight or sparse network often suffers from a significant performance drop and demands a high computational cost due to multiple rounds of pre-training; (2) existing compression methods on pre-trained speech models such as DistilHuBERT [8] are deficient under various resource constraints. Designing a general-purpose compression framework that enables different sizes of models under a tolerable training time is not well explored.

Inspired by the once-for-all approach [9], we propose LightHuBERT, an efficient model compression framework for lightweight and configurable speech pre-training, which consists of a once-for-all Transformer, a contextualized latent representation distillation objective, and a two-stage training strategy. Specifically, similar to [10], we build a weight-sharing Transformer supernet with channel-variable convolutional positional embeddings that can enable a Transformer architecture to scale in the embedding dimension, the head number, the feed-forward network (FFN) ratio, and the network depth. We introduce a pre-training distillation loss and perform masked self-supervised learning, where the student predicts contextualized representations within masked time steps that contain context information from a pre-trained teacher, the HuBERT BASE model [11]. A two-stage training strategy is designed to further improve the performance. Experiments on the automatic speech recognition and the SUPERB benchmark show that LightHuBERT attains superior results under various parameters.

The contributions of this paper are summarized as follows.

- We propose a lightweight and configurable model compression framework for speech pre-training to address the challenge of applying pre-trained models to various computational resources.

- We propose (1) a once-for-all Transformer with two individual supernets, (2) the contextualized latent representations to transfer knowledge, and (3) a two-stage training strategy to improve the performance.

- We demonstrate the effectiveness of the proposed method with pre-trained HuBERT BASE on automatic speech recognition task and the SUPERB benchmark.

2. Related Work

Large-scale pre-trained models such as wav2vec 2.0 [5], Hu-BERT [11], WavLM [12], SpeechT5 [1], and data2vec [13] have drawn much attention in the speech communities, due to

---
[†]Equal contribution. Corresponding author: Zhihua Wei (zhihua_wei@tongji.edu.cn).
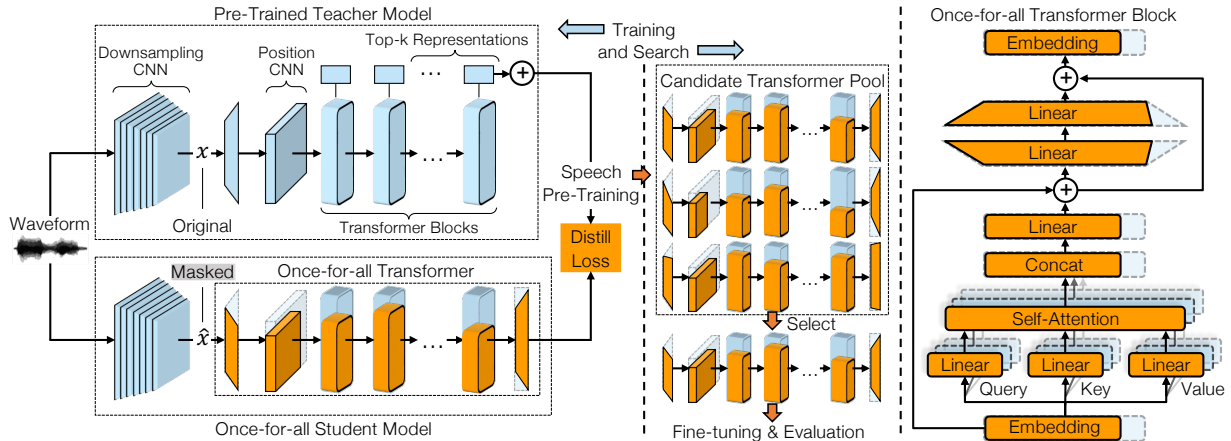
Figure 1: *Model compression for speech pre-training. All orange components are dynamic and share weights with large ones. Since this work focuses on pruning the encoder in speech pre-training, we keep the downsampling convolutional network static.* **Left:** *Given a pre-trained speech model, we distill a once-for-all Transformer containing all possible dimensions through weight sharing. We randomly sample different pairs of these variable dimensions at each training step so that one Transformer can support all architectures.* **Middle:** *We then select many sub-Transformers with different networks from the once-for-all Transformer, validate them without retraining, and finally select the best architecture given the parameter constraint to perform an evaluation.* **Right:** *The once-for-all Transformer block supports multiple dynamic dimensions: embedding dimension, attention dimension, head number, FFN ratio, and network depth.*

their excellent generalization capability across various speech applications and efficient use of large-scale unlabeled data. SU-PERB [14] has appeared as a popular benchmark to evaluate the performance of pre-trained models, where the pre-trained encoder is frozen and shared to investigate the capability of representations on various speech tasks, including phoneme recognition (PR), automatic speech recognition (ASR), key-word spotting (KS), query by example spoken term detection (QbE), speaker identification (SID), automatic speaker verification (ASV), speaker diarization (SD), intent classification (IC), slot filling (SF), and emotion recognition (ER).

Reducing architectural complexity of large-scale pre-trained models has become an indispensable research endeavor [15, 8, 16, 17, 18, 19, 20, 21]. DistilHuBERT [8] is proposed to distill hidden representations from HuBERT BASE. It creates a two-layer Transformer while retaining most performance on ten SUPERB tasks. Unlike the DistilHuBERT model, which creates a single compressed network, we distill a once-for-all network to deploy under various computational resources. PARP [17] is a magnitude-based unstructured pruning method that removes insignificant weights of wav2vec 2.0 based on a sparse network on low-resource ASR with the monolingual and cross-lingual transfer. Unlike PARP, we prune structured groups of weights, therefore, avoiding irregular sparse matrix operations that are hard to accelerate on hardware [22].

Once-for-all approaches [9] aim to create a weight-sharing supernet and obtain a huge number of architectures for efficient deployment with different resource constraints while maintaining the same level of accuracy as those trained independently [23, 24, 25, 26]. Examples can be found in image processing [9, 27], natural language processing [28], and speech processing [29]. Similarly, AutoFormer [10] utilizes weight entanglement to supernet training and enables Transformer blocks to share weights for their common in each layer. It allows a large number of subnets in the supernet to be as well-trained as ones trained from scratch. Unlike AutoFormer, we introduce a two-stage distillation to improve the supernet while utilizing a task-agnostic objective to learn speech representations.

3. LightHuBERT

As shown in Figure 1, in this section, we propose LightHu-BERT, a task-agnostic compression framework for reducing the model size of the Transformer encoder in speech pre-training. Specifically, we propose a once-for-all Transformer that enables automatic architecture search by pruning structured groups of weights. We transfer contextualized latent representations to sub-Transformers. A two-stage training strategy is proposed to improve the compressed models.

3.1. Once-for-All Transformer

Once-for-all Transformer refers to a Transformer architecture that contains various sub-Transformers, where different architectures share weights in a scaling manner. For example, a 256-dim linear layer is nested in a 512-dim linear layer. We design a once-for-all Transformer that contains five variable dimensions similar to AutoFormer [10]: embedding dimension, attention dimension, head number, FFN ratio, and network depth. We constrain the attention dimension (i.e., key, query, and value matrices) as the $64\times$ head number.

Since the interference between small and large networks degrades the performance of the large ones [29], we create two supernets to investigate models with significantly different model sizes, as shown in Table 1. Two supernets retain most Transformer blocks because existing Transformer-based speech pre-training usually applies deep networks to learn from large-scale unlabeled data and maintain superior performance, such as 12-layer HuBERT BASE and 24-layer HuBERT LARGE. To clarify, we denote a_{Largest}, a_{BASE}, and a_{SMALL} as shown in Table 2.

3.2. Pre-Training Distillation

We employ a masking-based pre-training distillation to transfer the knowledge of a pre-trained model. Specifically, we mask spans of latent speech representations in the student model and make the student model predict masked parts as the output of the teacher model. Inspired by [13], we introduce contextu-

Table 1: *Two supernets of once-for-all Transformers.*

	Small	Base
Embedding Dim	{256, 384, 512}	{512, 640, 768}
Head Number	{4, 6, 8}	{8, 10, 12}
FFN Ratio	{3.0, 3.5, 4.0}	{3.5, 4.0}
Network Depth	{10, 11, 12}	{12}
Parameters Range	11M – 45M	41M – 95M
Subnets Size	9.5×10^{11}	6.5×10^{9}

Table 2: *Frequently used items.*

Symbol	Description
a_{Largest}	The largest architecture as the HuBERT BASE size.
a_{BASE}	12-layer 640-embedding 10-head 2560-FFN subnet.
a_{SMALL}	12-layer 384-embedding 6-head 1536-FFN subnet.
OFA HuBERT	Once-for-all training initialized with the pre-trained HuBERT.
LightHuBERT	Two-stage training for a once-for-all HuBERT.

alized representations as the training target, i.e., average top-k normalized latent representations, where we set k = 8 as [13]. Unlike self-distillation in [13], we leverage a pre-trained speech model as the teacher. Formally, given a downsampled audio sequence x, the student is to minimize the L1 distance within masked time steps $\mathcal{M}$ as

$$\mathcal{L}\left(f^t(x), f^s(\hat{x})\right) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left| \bar{f}_i^t(x) - f_i^s(\hat{x}) \right|, \quad (1)$$

where $f^t(\cdot)$ denotes the teacher, $f^s(\cdot)$ denotes the student, $\hat{x}$ is the masked x with a masking probability of $p = 0.65$ as [13], and $\bar{f}_i^t(\cdot)$ denotes the training target at the i-th time step.

3.3. Two-Stage Training

To improve the performance of different weight-sharing architectures in the once-for-all Transformer, we propose a two-stage training strategy as follows.

- **Stage 1 - Distillation.** We train the largest architecture a_{Largest} of the once-for-all Transformer from scratch via the loss function of the pre-training distillation.

- **Stage 2 - Once-for-All Training.** We implement the once-for-all training on the supernet initialized by distilled weights. Specifically, we randomly sample a subnet from the supernet at each forwarding propagation during the supernet training.

The trained weights derived from Stage 1 serve as the initialization of Stage 2. Compared to existing pre-training objectives that force the top-layer representations to fit the targets, the pre-training distillation utilizes contextualized representations that provide receptive fields in different ranges and feature aggregation with various resolutions, which can be helpful to train subnets. For clarity, we define OFA HuBERT and LightHuBERT as shown in Table 2.

4. Experiments

We conduct our method with pre-trained HuBERT BASE. Pre-training models and the 10 hours ASR are conducted in fairseq [30]. The SUPERB tasks are implemented with S3PRL [14].

Table 3: *ASR results on $a_{Largest}$ between once-for-all Transformers and HuBERT. *HuBERT BASE is reproduced.*

Method	dev-clean	dev-other	test-clean	test-other
*HuBERT BASE [11]	9.6	16.3	9.6	16.9
OFA HuBERT	10.1	17.4	10.3	17.9
LightHuBERT	9.0	16.4	9.3	16.6
LightHuBERT a_{BASE}	9.4	17.3	9.6	17.5
LightHuBERT Stage 1	**8.6**	**14.5**	**8.7**	**14.8**

4.1. Experimental Setup

Model. The HuBERT BASE model has a 7-layer temporal convolution, 1-layer convolutional position, and a 12-layer Transformer encoder. The pre-trained weights are downloaded from [30]. Distilled student models have a similar architecture with an additional prediction head that predicts the training targets.

Datasets. For the pre-training task, we use a total of 960 hours of LibriSpeech audio [31]. For the ASR task, we fine-tune each model on the 10 hours labeled split of Librilight [32] and report the word error rate (WER) without a language model. For the SUPERB benchmark [14], we evaluate the models on ten tasks with the officially provided datasets, training recipes, and evaluation protocols, including PR, ASR, KS, QbE, SID, ASV, SD, IC, SF, and ER.

Pre-Training. The default pre-training is initialized with either the publicly released or the re-implemented weights. We pre-train once-for-all Transformers on 8 V100 GPUs with a batch size of around 119 seconds of audios per GPU for 200k steps, where the downsampling CNN is frozen to significantly reduce training time. Distilling a_{Largest} from scratch is implemented on 32 GPUs as the HuBERT BASE training recipe.

Search. We randomly search for subnets of the pre-trained models with the pre-trained distilled objective given parameters. For each pre-trained model, we search for 1,000 subnets. Besides, the minimal and maximal subnets are evaluated to estimate the potential performance bounds. Three architectures, i.e., a_{Largest}, a_{BASE}, and a_{SMALL}, are selected manually for evaluation on the SUPERB benchmark.

4.2. Automatic Speech Recognition

We conduct the once-for-all Transformer with HuBERT BASE on the ASR task and evaluate the largest architecture a_{Largest} from the base supernet. As shown in Table 3, the model of LightHuBERT at stage 1 outperforms the HuBERT teacher, which indicates the effectiveness of our distillation targets. LightHuBERT achieves superior performance than the OFA HuBERT, and its a_{BASE} achieves comparable performance to the HuBERT teacher with 29% reduced parameters, which suggests the two-stage strategy helps learn subnets. On the other hand, the performance of a_{Largest} slightly degrades in both OFA HuBERT and LightHuBERT after the once-for-all training, probably due to the interference between small and large networks caused by the once-for-all training.

To estimate the performance of the trained supernets, we evaluate some subnets from four trained once-for-all Transformers. These pre-trained Transformers include the OFA HuBERT and the LightHuBERT from small and base supernets. These subnets include the minimal, the maximal, two manually selected, and several randomly found architectures. The results are illustrated in Figure 2. The results of different selected architectures illustrate that these once-for-all Transformers obtain many well-trained sub-architectures. All chosen architectures

Table 4: *Speech pre-training evaluation on the SUPERB benchmark. Metrics include accuracy (Acc%), phoneme error rate (PER%), word error rate (WER%), maximum term weighted value (MTWV), F1 score (F1%), concept error rate (CER%), equal error rate (EER%), and diarization error rate (DER%). "Overall Score" denotes the average scores of all tasks as [12]. "Paral." denotes paralinguistics. "Distill." denotes distillation. ⋆HuBERT Teacher is reproduced and is slightly different from the released HuBERT.*

| Method | Params | Overall Score | Content | | | | Speaker | | | Semantics | | | Paral. |
			PR	ASR (WER)		KS	QbE	SID	ASV	SD	IC	SF		ER
			PER ↓	w/o ↓	w/ LM ↓	Acc ↑	MTWV ↑	Acc ↑	EER ↓	DER ↓	Acc ↑	F1 ↑	CER ↓	Acc ↑
HuBERT BASE [11]	95M	80.8	5.41	6.42	4.79	96.30	0.0736	81.42	5.11	5.88	98.34	88.53	25.20	64.92
DistilHuBERT [8]	23M	75.9	16.27	13.34	9.21	95.98	0.0511	73.54	8.55	6.19	94.99	82.57	35.59	63.02
⋆HuBERT Teacher	95M	81.1	5.23	6.62	4.96	96.66	**0.0879**	**82.90**	**4.94**	**5.45**	98.21	**88.81**	**25.30**	64.68
Stage 1 Distill. $a_{Largest}$	95M	81.0	**4.15**	**5.71**	**4.20**	**96.82**	0.0737	80.01	5.14	5.51	**98.50**	88.44	25.92	**66.25**
LightHuBERT $a_{Largest}$	95M	80.4	4.56	6.43	4.80	96.40	0.0642	77.58	5.43	5.85	98.21	88.78	25.32	64.93
LightHuBERT a_{Base}	68M	80.4	4.71	6.72	4.97	95.75	0.0767	77.24	5.55	5.73	98.00	88.79	26.06	65.55
LightHuBERT a_{Small}	27M	79.1	6.60	8.33	6.04	96.07	0.0764	69.70	5.42	5.85	98.23	87.58	26.90	64.12
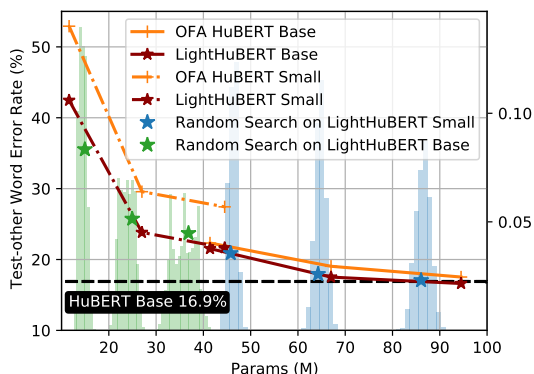
Figure 2: *ASR results of the test-other between LightHuBERT and OFA HuBERT. Green and blue histograms denote the distributions of parameters of subnets sampled from the small and base supernets, respectively. We evaluate six subnets found by the random search in the LightHuBERT. Three of them are found given 15M, 25M, and 37M parameters in the small supernet, and the others are found given 47M, 65M, and 87M parameters in the base supernet.*

from LightHuBERT outperform the networks from OFA Hu-BERT, which represents the superiority of the two-stage training strategy.

4.3. Universal Representation Evaluation

To provide a comprehensive testbed for the generalizability of our compressed pre-trained models, we evaluate four models derived from LightHuBERT on the SUPERB benchmark. As shown in Table 4, we can draw the following conclusions: (1) The proposed LightHuBERT can create compressed models that retain comparable performance on the SUPERB tasks. The a_{SMALL} with around 28% parameter achieves only a 2 point drop in overall score, which significantly outperforms Distil-HuBERT by an absolute improvement of 3.2 point. (2) The LightHuBERT at Stage 1 achieves superior performance than the HuBERT teacher in five tasks while maintaining comparable performance in other tasks, which demonstrates the effectiveness of the proposed pre-training distillation to learn universal speech representations. (3) We find the a_{SMALL} achieves 3.5× compression ratio in KS, ASV, and IC tasks with a slight accuracy loss, and the a_{BASE} attains comparable performance with the HuBERT teacher by saving 29% parameters in most tasks such as PR and ASR, which indicates that the model size could vary across SUPERB tasks.

Table 5: *Ablation study with a_{BASE} from the base supernet. We report WER (%) on the ASR task without a language model.*

Method	test-clean ↓	test-other ↓
LightHuBERT	9.6	17.5
- Stage 1 Distillation	13.3 (+3.7)	22.4 (+4.9)
- Stage 2 Once-for-All Training	12.2 (+2.5)	22.8 (+5.3)

4.4. Ablation Study

As shown in Table 5, an ablation study is conducted to investigate the pre-training distillation and the once-for-all training by evaluating the a_{BASE} on the ASR task. The performance of the pre-trained a_{BASE} drops consistently without either pre-training distillation or once-for-all training, which suggests that both pre-training distillation and once-for-all contribute to improving the supernet. Both stages significantly enhance the performance of the subnet a_{BASE} by over 20% relative improvement in WER, which reveals that initializing weights and sampling diverse subnets during training help obtain well-trained networks.

5. Conclusion

In this paper, we propose LightHuBERT, a once-for-all Transformer compression framework to produce many pre-trained models in different sizes, making it available to run pre-trained models under various computational resources. Experiments with the pre-trained HuBERT BASE on the 10 hour ASR task and the SUPERB benchmark demonstrate the effectiveness of the proposed LightHuBERT. Considering different parameters, the distilled model achieves superior performance compared to the HuBERT BASE teacher in most speech tasks with the size of the teacher model. LightHuBERT obtains comparable performance to the HuBERT teacher in most tasks with 29% reduced parameters. We achieve a 3.5× compression ratio in ASV, KS, and IC tasks with a slight accuracy loss while outperforming DistilHuBERT by an absolute improvement of 3.2 point in terms of the overall score. In future work, we will jointly employ parameter compression [33] and task-dependent model compression [18] to further compress the pre-trained model.

6. Acknowledgements

7. References

[1] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

[2] Y.-A. Chung and J. Glass, "Speech2Vec: A sequence-to-sequence framework for learning word embeddings from speech," in *Proceedings of the Interspeech*, 2018, pp. 811–815.

[3] Y.-S. Chuang, C.-L. Liu, H.-Y. Lee, and L.-s. Lee, "SpeechBERT: An audio-and-text jointly learned language model for end-to-end spoken question answering," in *Proceedings of the Interspeech*, 2020, pp. 4168–4172.

[4] X. Song, G. Wang, Y. Huang, Z. Wu, D. Su, and H. Meng, "Speech-XLNet: Unsupervised acoustic model pretraining for self-attention networks," in *Proceedings of the Interspeech*, 2020.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[6] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "UniSpeech: Unified speech representation learning with labeled and unlabeled data," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 10 937–10 947.

[7] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *arXiv preprint arXiv:2108.06209*, 2021.

[8] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[9] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once for All: Train one network and specialize it for efficient deployment," in *International Conference on Learning Representations*, 2020.

[10] M. Chen, H. Peng, J. Fu, and H. Ling, "AutoFormer: Searching transformers for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 270–12 280.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[12] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. W. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[13] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint arXiv:2202.03555*, 2022.

[14] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech processing universal performance benchmark," in *Proceedings of the Interspeech*, 2021, pp. 1194–1198.

[15] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - Advances in Neural Information Processing Systems*, 2019.

[16] G. Aguilar, Y. Ling, Y. Zhang, B. Yao, X. Fan, and C. Guo, "Knowledge distillation from internal representations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7350–7357.

[17] C.-I. J. Lai, Y. Zhang, A. H. Liu, S. Chang, Y.-L. Liao, Y.-S. Chuang, K. Qian, S. Khurana, D. Cox, and J. Glass, "PARP: Prune, adjust and re-prune for self-supervised speech recognition," in *Proceedings of the 35th Advances in Neural Information Processing Systems*, vol. 34, 2021.

[18] S. Zhang, X. Zheng, C. Yang, Y. Li, Y. Wang, F. Chao, M. Wang, S. Li, J. Yang, and R. Ji, "You Only Compress Once: Towards effective and elastic BERT compression via exploit-explore stochastic nature gradient," *arXiv preprint arXiv:2106.02435*, 2021.

[19] S. Yu, T. Chen, J. Shen, H. Yuan, J. Tan, S. Yang, J. Liu, and Z. Wang, "Unified visual transformer compression," in *Proceedings of the International Conference on Learning Representations*, 2021.

[20] O. Zafrir, A. Larey, G. Boudoukh, H. Shen, and M. Wasserblat, "Prune Once for All: Sparse pre-trained language models," in *Proceedings of the 35th Conference on Neural Information Processing Systems*, 2021.

[21] Z. Peng, A. Budhkar, I. Tuil, J. Levy, P. Sobhani, R. Cohen, and J. Nassour, "Shrinking Bigfoot: Reducing wav2vec 2.0 footprint," in *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, 2021, pp. 134–141.

[22] T. Gale, E. Elsen, and S. Hooker, "The state of sparsity in deep neural networks," in *Proceedings of the Joint Workshop on On-Device Machine Learning and Compact Deep Neural Network Representations*, 2019.

[23] L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485–532, 2020.

[24] X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, p. 106622, 2021.

[25] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 712–733, 2020.

[26] Z. Lu, G. Sreekumar, E. Goodman, W. Banzhaf, K. Deb, and V. N. Boddeti, "Neural architecture transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 9, pp. 2971–2989, 2021.

[27] M. Li, J. Lin, Y. Ding, Z. Liu, J.-Y. Zhu, and S. Han, "GAN Compression: Efficient architectures for interactive conditional GANs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5283–5293.

[28] L. Hou, Z. Huang, L. Shang, X. Jiang, X. Chen, and Q. Liu, "DynaBERT: Dynamic BERT with adaptive width and depth," in *Proceedings of the 34th Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9782–9793.

[29] R. Wang, Z. Wei, H. Duan, S. Ji, Y. Long, and Z. Hong, "EfficientTDNN: Efficient architecture search for speaker recognition," *arXiv preprint arXiv:2103.13581*, 2021.

[30] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2019, pp. 48–53.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[32] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-Light: A benchmark for ASR with limited or no supervision," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7669–7673.

[33] B. Wang, Y. Ren, L. Shang, X. Jiang, and Q. Liu, "Exploring extreme parameter compression for pre-trained language models," in *Proceedings of the International Conference on Learning Representations*, 2021.