Ensemble Knowledge Distillation from Speech SSL Models Considering Inter-teacher Differences

Pei-Jun Liao[1,2], *Hung-Yi Lee*[1], *Hsin-Min Wang*[2]

[1]College of Electrical Engineering and Computer Science, National Taiwan University, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan

d06921018@ntu.edu.tw, hungyilee@ntu.edu.tw, whm@iis.sinica.edu.tw

Abstract

In the realm of speech processing, Self-Supervised Learning (SSL) models such as HuBERT are widely used in various speech tasks such as Automatic Speech Recognition (ASR) and Spoken Language Understanding (SLU), and have achieved impressive results. However, these speech SSL models are often large and require significant computational resources. Many previous studies have used Knowledge Distillation (KD) to learn compact models from complex models. Ensemble Knowledge Distillation (EKD) transfers multiple SSL models into a single student model via multiple prediction heads. We focus on the differences among teacher models and use them as residuals in student model learning to achieve additional learning goals. Residual Prediction Heads and Residual Regularization are proposed. Combining RobustHuBERT with WavLM+ or Data2vec-base as teachers, we evaluate the resulting student models on 6 tasks in the SUPERB benchmark. The results show that taking residuals into account improves performance.

Index Terms: Self-Supervised Learning, Knowledge Distillation, Spoken Language Understanding, Emotion Recognition

1. Introduction

In recent years, the concept of foundation models, starting with BERT [1], has profoundly influenced problem-solving approaches. The key advantage of foundation models lies in their utilization of large amounts of unlabeled data for training the model through Self-Supervised Learning (SSL), referred to as pre-training or the upstream task. The parameters are then fine-tuned in so-called downstream tasks, constituting the second step of training. Foundation models have demonstrated their versatility across multiple tasks, including good performance on benchmarks such as the GLUE benchmark [2] and Reading Comprehension Question Answering in the Natural Language Processing (NLP) domain. Leveraging the success observed in NLP, speech processing has also adopted foundation models to address tasks within Spoken Language Understanding (SLU) [3], such as Sentiment Analysis (SA) [4], Automatic Speech Recognition (ASR) [5], and Spoken Question Answering (SQA) [6].

Despite the effectiveness of foundation models in various application tasks, their reliance on the transformer-based architecture [7] involving self-attention computation poses challenges due to the O(N^2) complexity, where N is the encoding sequence length. Additionally, original SSL models often comprise numerous layers. DistilBERT [8] addresses this issue by employing Knowledge Distillation (KD) [9] to compress the BERT model into a smaller version while retaining the language understanding ability. Building upon this concept, DistilHuBERT [10] compresses the HuBERT model [11] to 2 transformer layers instead of the original 12 transformer layers.

Inspired by the idea of combining multiple teacher models to enhance the performance of distilled models in ASR [12], Ensemble Knowledge Distillation (EKD) [13] takes a similar approach to combine 3 speech SSL models to improve DistilHuBERT. Based on EKD, we believe that learning the differences of multiple teacher models may improve performance in downstream speech tasks. We employ two approaches, namely Residual Prediction Heads and Residual Regularization, to enhance the EKD speech model. Our improvements are demonstrated on 3 teacher SSL models (WavLM+ [14], Data2vec-base [15], and RobustHuBERT [16]) on the SUPERB benchmark [17].

Our contributions are:

1. We introduce Residual Learning into EKD by leveraging the residual values of different teacher SSL models and demonstrate performance improvements on the SUPERB benchmark.
2. We apply KD to Data2vec-base, which has not been done in previous work.

In this study, we only applied EKD to combinations of two of the three teacher models. We plan to study the combination of all three models and their combinations with other speech SSL models in future research.

2. Related work

Based on the success of SSL language models such as BERT-like models [18, 19], speech SSL models have been introduced in the field of speech processing. wav2vec [20], one of the pioneers of speech SSL models, and its extended versions wav2vec2 [21], W2V-BERT [22], and vq-wav2vec [23] are widely used in speech processing research. Building upon the concept of wav2vec, HuBERT introduces multi-stage quantization as the target of masking learning. WavLM further enhances performance by incorporating data augmentation such as noise addition and modified position bias, resulting in state-of-the-art results on many tasks in the SUPERB benchmark.

Considering that the large number of parameters of speech SSL models leads to increased computational costs, several techniques have emerged to compress the model size while retaining the knowledge capacity. Modeling after ALBERT [24], which utilizes weight-sharing and factorization to revamp BERT to save computing resources, Audio ALBERT [25] is also introduced. In addition, DPHuBERT [26] utilizes joint distillation and pruning techniques to reduce model size, while DistilHuBERT employs KD techniques to address this challenge.

In the domain of text processing, EKD from multiple teacher models have demonstrated superior performance compared to KD from single teacher models in certain tasks [27, 28,

29]. EKD from multiple speech SSL models has also shown success on some hidden sets in SUPERB. Inspired by the residual utilization in Vector Quantization (VQ) [30, 31], we adopt the residuals as learning targets in addition to the teachers' layer outputs. Building upon EKD, we integrate residual prediction into the KD learning step to further enhance performance.

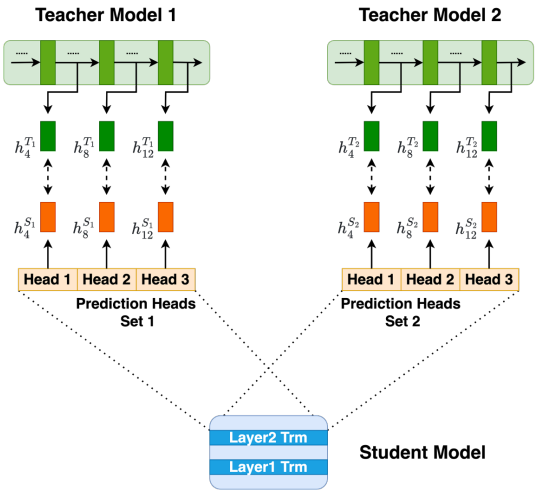


Figure 1: *Ensemble knowledge distillation.*

3. Methodology

3.1. Knowledge distillation

In the KD process of DistilHuBERT, the representations of the 4th, 8th, and 12th layers of the HuBERT Base model are utilized as the targets provided by the teacher model. L1 and cosine similarity losses are employed to enable DistilHuBERT's main model (only two transformer encoder layers) and three additional prediction heads to learn the targets. The loss function for KD of DistilHuBERT is

$$KD(h^T, h^S) = \frac{1}{3} \sum_{i \in \{4,8,12\}} KD_{layer}(h_i^T, h_i^S), \quad (1)$$

where h denotes the sequence representation, i is the layer index, and T and S respectively indicate the student model and the teacher model, and $KD_{layer}(h_i^T, h_i^S)$ is formulated as

$$KD_{layer}(h_i^T, h_i^S) = \|h_i^T - h_i^S\|_1 - \log \sigma(cossim(h_i^T, h_i^S)), \quad (2)$$

where $cossim$ denotes the cosine similarity, and σ denotes the sigmoid function. The same KD process can also be applied to RobustHuBERT, WavLM+ and Data2vec-base to obtain the corresponding compressed models.

3.2. Ensemble knowledge distillation

The EKD process proposed in [13] to enhance DistilHuBERT by combining multiple teacher models is shown in Figure 1. The loss function is

$$EKD(h^T, h^S) = \frac{1}{M} \sum_{m=1}^{M} KD(h^{T_m}, h^{S_m}), \quad (3)$$

where m is the teacher index, and M is the number of teachers.

In [13], the authors tried a variety of KD methods, including using a single set of prediction heads to correspond to the average or concatenation of each layer of multiple teachers, and using multiple sets of prediction heads to correspond to each layer of each teacher (see Figure 1). The approach utilizing multiple sets of prediction heads produced the best results.

3.3. Residual ensemble knowledge distillation

In this work, we present an enhancement rooted in the EKD method. In the KD stage of the upstream model, we advocate a technique that takes into account the information differences among teacher models. Termed Residual Ensemble Knowledge Distillation, we introduce two methods in it: Residual Prediction Heads and Residual Regularization.

3.3.1. Residual prediction heads

As shown in Figure 2(a), we use an additional set of prediction heads designed specifically to predict the difference values between the target layers of two teachers. This enables the model to discern differences between SSL models, thereby enhancing the efficacy of the distiller in the ensemble configuration. In the case of multiple teachers, the loss function is

$$REKD_{RP}(h^T, h^S) = EKD(h^T, h^S) + \frac{1}{\binom{M}{2}} \sum_{m=1}^{M} \sum_{n=1, n \neq m}^{M} KD([h^{T_m} - h^{T_n}], h^{S_{m,n}}), \quad (4)$$

where $S_{m,n}$ denotes the set of residual prediction heads corresponding to the pair of teacher n and teacher m.

3.3.2. Residual regularization

As shown in Figure 2(b), we use the difference values between the two teachers as regularization terms for the two sets of prediction heads in the student model corresponding to the two teachers. In the case of multiple teachers, the loss function is

$$REKD_{RR}(h^T, h^S) = EKD(h^T, h^S) + Reg, \quad (5)$$

where Reg is calculated as

$$\frac{1}{\binom{M}{2}} \sum_{m=1}^{M} \sum_{n=1, n \neq m}^{M} KD([h^{T_m} - h^{T_n}], [h^{S_m} - h^{S_n}]). \quad (6)$$

4. Experimental setup

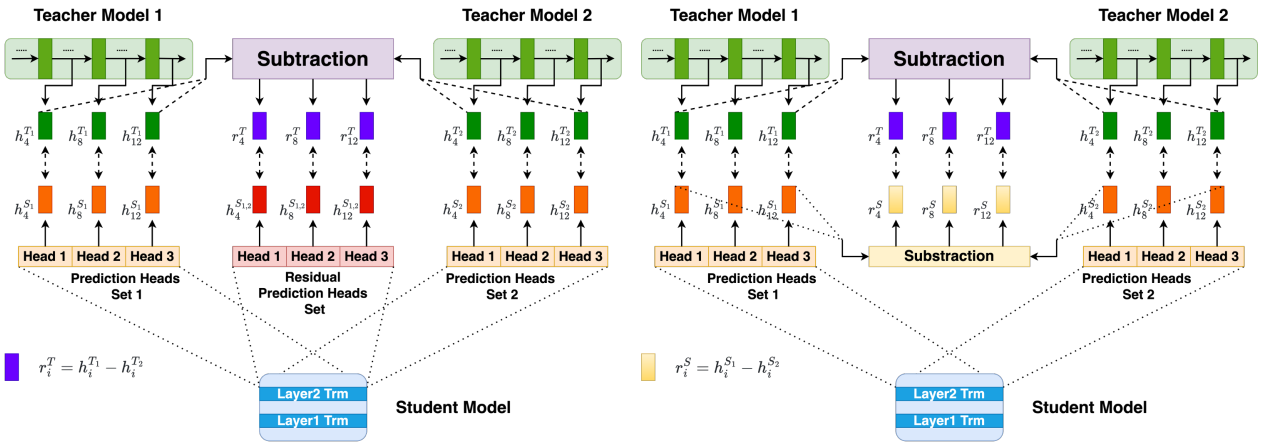
4.1. SUPERB benchmark

We conduct experiments on the Speech processing Universal PERformance Benchmark (SUPERB). Specifically, we focus on 6 sub-tasks in this benchmark: Phone Recognition (PR), Slot Filling (SF), Emotion Recognition (ER), Intent Classification (IC), Keyword Spotting (KS), and Low resource ASR (10 hours of training).

4.1.1. Intent classification, keyword spotting, and emotion recognition

The IC, KS, and ER tasks are classified as audio sequence classification tasks in SUPERB.

There are two datasets for SUPERB's KS and IC tasks: Speech Commands [32] and Fluent Speech Commands [33]. The Speech Commands dataset focuses on classifying single spoken words with background noise, while the Fluent Speech



(a) *EKD with Residual Prediction Heads* (b) *EKD with Residual Regularization*

Figure 2: *Residual ensemble knowledge distillation.*

Command dataset involves a three-fold intent-classification task designed for smart home applications, including sub-tasks related to Location, Object, and Action. Both tasks are evaluated in terms of accuracy.

For the ER task in SUPERB, we use the IEMOCAP dataset [34], which includes audio data and corresponding face videos. However, only audio is considered for evaluation in SUPERB. The IEMOCAP dataset comprises 5 sessions, each with its own training and test sets. Performance is evaluated based on the average accuracy over 5 sessions.

4.1.2. Slot filling, phone recognition, and ASR

The SF, PR and ASR tasks are categorized as sequence generation tasks in SUPERB.

The dataset for the SF task is an audio adaptation of the text-based Snips dataset [35, 36], known as Audio Snips. For the Audio Snips SF task, a sequence-to-sequence generation method is used to convert speech sequences into text sequences and identify slot types and entity words. Metrics such as slot-type-F1 score and slot-value-CER are used to evaluate model performance.

The dataset for the PR task comes from Librispeech [37], specifically the 100 hours of Librispeech-clean. Phoneme labels are generated from Ground Truth text and G2P conversion. The PR task is a simplified task of ASR that focuses on recognizing phone-level symbols and is evaluated using the phone error rate (PER).

The dataset for the low resource ASR task is Libri-light 10 hours [38], derived from the LibriVox [1] project. It includes a 10-hour training set, as well as development and test sets from Librispeech. The task is evaluated in terms of word error rate (WER) and character error rate (CER) under clean and other test conditions.

4.2. Teacher models

Base on the results of [13], we also select RobustHuBERT as one of the teacher models. In this study, we combine RobustHuBERT with WavLM+ or Data2vec-base.

RobustHuBERT is an enhanced version of HuBERT implemented by integrating Domain-Adaptive Pre-Training (DAPT) into the HuBERT-base model. Training data from Librispeech is augmented with noise distortion, pitch-shift, reverberation, and band-rejection techniques.

WavLM+ is the leading variant on the SUPERB leaderboard among various WavLM models. It is very similar in structure to HuBERT, but incorporates gated-relative position bias in the transformer-based model and employs noise data augmentation.

Data2vec-base is a versatile model that integrates text, image, and audio modalities. During audio version pre-training, audio frame spans are masked, and predictions are made accordingly.

4.3. Training details

We follow the evaluation procedure outlined in s3prl[2] to fix the SSL model while training the downstream task model. For the ASR task, we adjust the training steps to 50,000. We adopt the DistilHuBERT recipe from s3prl to the KD of upstream models, setting the batch size to 16 and training for 300,000 steps with a learning rate of 2e-4. Each distillation model is trained individually on a single NVIDIA-RTX-3090 GPU, taking approximately 60 hours to complete.

5. Results

The performance of different distillation models evaluated on 6 tasks of SUPERB is shown in Table 1. HB refers to DistillHuBERT [10], while DV, WL, and RHB refer to distillation models from Data2vec-base, WavLM+, and RobustHuBERT, respectively, using the same KD technique as DistillHuBERT. For HB, the results are from the SUPERB leaderboard. WL-RHB denotes the model distilled from WavLM+ and RobustHuBERT, while DV-RHB denotes the model distilled from Data2vec-base and RobustHuBERT. The WL-RHB and DV-RHB models without subscripts means that they are trained through the ensemble knowledge distillation method proposed in [13]. The WL-RHB and DV-RHB models with RP subscript imply the use of Residual Prediction Heads in the proposed residual ensemble knowledge distillation method, while the models with RR subscript imply the use of Residual Regularization in the proposed

[1] https://librivox.org

[2] Recipes for upstream and downstream are in https://github.com/s3prl/s3prl

Table 1: *Performance of different distillation models evaluated on 6 tasks of SUPERB. ↑ indicates that higher values are better, while ↓ indicates that lower values are better. The best performance in each task is shown in red. If our proposed model outperforms the corresponding baseline distillation model from multiple teachers, the performance will be highlighted in gray. If our proposed model outperforms the corresponding baseline distillation model from a single teacher, the performance is shown in bold.*

	IC	KS	ER	SF		PR	ASR			
							clean		other	
	Acc (↑)	Acc (↑)	Acc (↑)	F1 (↑)	CER (↓)	PER (↓)	WER (↓)	CER (↓)	WER (↓)	CER (↓)
HB	94.99	95.98	63.02	82.57	35.59	16.27	-	-	-	-
DV	93.62	95.75	62.75	85.30	31.17	10.79	21.56	6.97	41.36	16.62
WL	95.89	96.72	63.61	83.48	33.96	13.49	25.91	8.19	45.38	17.61
RHB	94.41	96.27	63.41	83.42	33.51	12.89	25.83	8.27	45.67	17.93
WL-RHB	94.36	96.27	63.91	83.81	32.07	12.47	24.17	7.72	43.38	16.91
WL-RHB$_{RP}$	95.41	96.62	62.78	**83.85**	**32.19**	**12.15**	**24.07**	**7.68**	**42.89**	**16.66**
WL-RHB$_{RR}$	95.41	96.11	62.91	82.08	34.08	**12.35**	**24.91**	**7.91**	**44.10**	**17.21**
DV-RHB	93.91	96.11	64.35	85.47	31.37	10.87	21.85	6.97	40.48	15.91
DV-RHB$_{RP}$	**94.91**	96.07	**63.78**	**85.92**	**30.75**	**10.76**	21.90	**6.96**	42.21	**16.51**
DV-RHB$_{RR}$	94.07	95.78	62.65	84.69	31.70	11.66	22.95	7.31	42.34	16.63

residual ensemble knowledge distillation method. All distillation models have the same structure and 23 million parameters.

5.1. Single teacher knowledge distillation baselines

From the upper half of Table 1, it is worth noting that distilled WavLM+ (WL) outperforms distilled RobustHuBERT (RHB) and distilled Data2vec-base (DV) in the three sequence classification tasks of IC, KS, and ER. However, in the three sequence generation tasks of SF, PR, and ASR, DV performs best among the four single-teacher distillation models.

5.2. Multiple teachers knowledge distillation baselines

From Table 1, we can see that both WL-RHB and DV-RHB do not always outperform their corresponding single-teacher distillation models. Furthermore, DV-RHB outperforms WL-RHB in the ER task and three sequence generation tasks of SF, PR, and ASR. The results show that the complementarity between Data2vec-base and RobustHuBERT is higher than that between WavLM+ and RobustHuBERT.

5.3. Multiple teachers knowledge distillation with residuals

From Table 1 we can see that WL-RHB$_{RP}$ outperforms WL-RHB on 8 out of 10 evaluation metrics on 6 tasks. But WL-RHB$_{RR}$ performs worse than WL-RHB in most cases. Compared with WL and RHB, WL-RHB$_{RP}$ outperforms both in the three sequence generation tasks of SF, PR, and ASR, but WL-RHB$_{RR}$ only outperforms both in the two tasks of PR and ASR. The results demonstrate the effectiveness of using Residual Prediction Heads in the proposed residual ensemble knowledge distillation method.

When using Data2vec-base and RobustHuBERT as teacher models, DV-RHB$_{RP}$ outperforms DV-RHB in 3 out of 6 tasks (i.e., IC, SF, and PR). DV-RHB$_{RP}$ and DV-RHB perform comparably. But DV-RHB$_{RR}$ only outperforms DV-RHB in the IC task. DV-RHB$_{RR}$ performs worse than DV-RHB. Furthermore, compared with DV and RHB, DV-RHB$_{RP}$ outperforms both on 7 out of 10 evaluation metrics on 6 tasks. The results again verify the effectiveness of using Residual Prediction Heads in the proposed residual ensemble knowledge distillation method. The results also imply that DV-RHB$_{RP}$ is particularly useful in relatively resource-rich tasks such as IC, SF, and PR.

5.4. Discussion about DV-RHB on ASR

We observe a significant performance gap between the single-teacher distillation models DV and RHB on the ASR task (e.g., under the clean test condition, their WERs are 21.56% and 25.83%, respectively). This may be why DV-RHB$_{RP}$ performs worse than DV on this task. Therefore, the process of ensemble knowledge distillation may require modifications, such as implementing weighting mechanisms for teachers. This will be included in our future work.

6. Conclusion

In this study, we have improved the ensemble knowledge distillation of multiple teacher models. By leveraging multiple prediction heads, we propose utilizing residuals between teachers as learning targets during knowledge distillation. This enables the student model to learn differences among teachers, thereby extending its capabilities. We implemented this approach across 6 SUPERB tasks and achieved improvements. In our future work, we plan to explore more combinations of speech SSL models and evaluate their performance on classic speech tasks such as rich-resource ASR and Speech Translation. We anticipate that the introduction of residuals will open up new avenues for knowledge distillation in the field of speech processing.

7. Acknowledgements

We thank Ming-Chi Yen for maintaining the GPU server infrastructure. Additionally, thanks to Jeff Wang for providing access to the IEMOCAP, Speech Command, and Fluent Speech Command datasets. We would also like to thank Shu-Wen Yang for maintaining the s3prl platform and Chia-Hua Wu for providing valuable insights and guidance on experimental reports.

8. References

[1] J. Devlin, M.-W. Chang, L. Kenton, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.

[2] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. ICLR*, 2019.

[3] Y. Peng, S. Arora, Y. Higuchi, Y. Ueda, S. Kumar, K. Ganesan, S. Dalmia, X. Chang, and S. Watanabe, "A study on the integra-

tion of pre-trained SSL, ASR, LM and SLU models for spoken language understanding," in *Proc. SLT*, 2022.

[4] S. Shon, A. Pasad, F. Wu, P. Brusco, Y. Artzi, K. Livescu, and K. J. Han, "SLUE: New benchmark tasks for spoken language understanding evaluation on natural speech," in *Proc. ICASSP*, 2022.

[5] F. Wu, K. Kim, S. Watanabe, K. J. Han, R. McDonald, K. Q. Weinberger, and Y. Artzi, "Wav2Seq: Pre-training speech-to-text encoder-decoder models using pseudo languages," in *Proc. ICASSP*, 2023.

[6] G.-T. Lin, Y.-S. Chuang, H.-L. Chung, S.-w. Yang, H.-J. Chen, S. A. Dong, S.-W. Li, A. Mohamed, H.-y. Lee, and L.-s. Lee, "DUAL: Discrete spoken unit adaptive learning for textless spoken question answering," in *Proc. Interspeech*, 2022.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[9] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[10] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT," in *Proc. ICASSP*, 2022.

[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] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in *Proc. Interspeech*, 2016.

[13] K.-P. Huang, T.-h. Feng, Y.-K. Fu, T.-Y. Hsu, P.-C. Yen, W.-C. Tseng, K.-W. Chang, and H.-y. Lee, "Ensemble knowledge distillation of self-supervised speech models," in *Proc. ICASSP*, 2023.

[14] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, and X. Xiao, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[15] 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," in *Proc. ICML*, 2022.

[16] K.-P. Huang, Y.-K. Fu, T.-Y. Hsu, F. R. Gutierrez, F.-L. Wang, L.-H. Tseng, Y. Zhang, and H.-y. Lee, "Improving generalizability of distilled self-supervised speech processing models under distorted settings," in *Proc. SLT*, 2022.

[17] S.-w. 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.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021.

[18] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. NeurIPS*, 2019.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa word embedding based power grid dispatching entity recognition," in *Proc. IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, 2020.

[20] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019.

[21] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[22] 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," in *Proc. ASRU*, 2021.

[23] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *Proc. ICLR*, 2020.

[24] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. ICLR*, 2020.

[25] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, Y.-H. Chen, S.-W. Li, and H.-y. Lee, "Audio ALBERT: A lite BERT for self-supervised learning of audio representation," in *Proc. SLT*, 2021.

[26] Y. Peng, Y. Sudo, S. Muhammad, and S. Watanabe, "DPHuBERT: Joint distillation and pruning of self-supervised speech models," in *Proc. Interspeech*, 2023.

[27] Y. Gao, T. Parcollet, and N. D. Lane, "Distilling knowledge from ensembles of acoustic models for joint CTC-attention end-to-end speech recognition," in *Proc. ASRU*, 2021.

[28] C. Wu, F. Wu, and Y. Huang, "One teacher is enough? pre-trained language model distillation from multiple teachers," in *Proc. ACL-IJCNLP*, 2021.

[29] C. Wu, F. Wu, T. Qi, and Y. Huang, "Unified and effective ensemble knowledge distillation," *arXiv preprint arXiv:2204.00548*, 2022.

[30] Y.-C. Wu, I. D. Gebru, D. Marković, and A. Richard, "Audiodec: An open-source streaming high-fidelity neural audio codec," in *Proc. ICASSP*, 2023.

[31] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[32] P. Warden, "Speech Commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[33] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Proc. Interspeech*, 2019.

[34] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[35] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, and T. Lavril, "Snips Voice Platform: An embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[36] C.-I. Lai, Y.-S. Chuang, H.-Y. Lee, S.-W. Li, and J. Glass, "Semi-supervised spoken language understanding via self-supervised speech and language model pretraining," in *Proc. ICASSP*, 2021.

[37] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[38] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, and C. Fuegen, "Libri-light: A benchmark for ASR with limited or no supervision," in *Proc. ICASSP*, 2020.