ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095591

DATA2VEC-SG: IMPROVING SELF-SUPERVISED LEARNING REPRESENTATIONS FOR SPEECH GENERATION TASKS

Heming Wang[1*], *Yao Qian*[2], *Hemin Yang*[2], *Nauyuki Kanda*[2], *Peidong Wang*[2], *Takuya Yoshioka*[2], *Xiaofei Wang*[2], *Yiming Wang*[2], *Shujie Liu*[2], *Zhuo Chen*[2], *DeLiang Wang*[1], *Michael Zeng*[2]

[1]The Ohio State University, USA [2]Microsoft Corporation, USA

wang.11401@osu.edu, {yaoqian,heyang,nakanda,peidongwang,tayoshio,xiaofewa}@microsoft.com,
{yimingwang,shujliu,zhuc}@microsoft.com,dwang@cse.ohio-state.edu,nzeng@microsoft.com

ABSTRACT

Self-supervised learning has been successfully applied to various speech recognition and understanding tasks. However, for generative tasks such as speech enhancement and speech separation, most self-supervised speech representations did not show substantial improvements. To deal with this problem, in this paper, we propose data2vec-SG (Speech Generation), which is a teacher-student learning framework that addresses speech generation tasks. Our data2vec-SG introduces a reconstruction module into data2vec [1] and enforces the representations to contain not only the semantic information but also the acoustic knowledge to generate clean speech waveforms. Experimental results demonstrate that the proposed framework boosts the performance of various speech generation tasks including speech enhancement, speech separation, and packet loss concealment. Meanwhile, the learned representation is also capable of helping other downstream tasks, which is demonstrated by the good performance in the speech recognition task in both clean and noisy conditions.

Index Terms— self-supervised learning, speech enhancement, speech separation, packet loss concealment, automatic speech recognition

1. INTRODUCTION

Self-supervised learning (SSL) has achieved massive success across many research fields including language, vision, and speech. It leverages a large amount of data without human annotations and learns a universal representation that is beneficial for various downstream tasks. In the field of speech, existing SSL methods can be categorized into several groups, including auto-regressive methods like APC [2], generative tasks like TERA [3]. Also, discriminative methods like wav2vec2.0 [4], Hubert [5], Unispeech-SAT [6] and WavLM [7]. Specifically, by leveraging large-scale pre-training, we expect to obtain a representation that is more effective for downstream training compared with adopting waveform or spectrogram features directly extracted from the input. For the automatic speech recognition (ASR) task, with a very limited amount of labeled data (10 minutes), SSL-based representations can still achieve remarkable word error rate (WER) performance.

However, even with the significant performance improvement for numerous tasks, SSL-based representations do not exert a clear advantage for generative tasks such as speech separation and speech enhancement. On the Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities (SUPERB-SG) [8], compared with the simple mel-filterbank (FBANK) features, many SSL representations even perform worse in terms of both short-time objective intelligibility (STOI) [9] and perceptual evaluation of speech quality (PESQ) [10]. For a few advanced SSL models, the benefit of employing pre-trained representations is marginal. This relative improvement is much smaller than the performance gain we observed in other non-generative tasks. The purpose of this study is to fill this gap and explores improving the generative capability of existing SSL methods.

There have been studies that aim to improve the speech enhancement performance of pre-trained speech representations. Hang et al. [11] first systematically evaluated the effectiveness of 13 SSL representations for speech enhancement and separation. Through a denoising mask prediction task, they conclude that only some of tested SSL representations (wav2vec2 [4], Unispeech-SAT [6], Hubert [5], WavLM [7]) outperform the FBANK features, and the authors claim that the information that is required for clean speech reconstruction might be lost in deeper layers of SSL models. Hung et al. [12] proposed to include incorporating the logarithm spectrogram along with the SSL representations and fine-tuning the upstream model with the enhancement network to compensate for the information loss of the learned representations. Kataria et al. [13] compared representation from various models for speech enhancement and observed a better performance from classifier-based embedding (acoustic event classifier) over SSL models [4, 14].

In this paper, we propose data2vec-SG, an SSL framework based on data2vec [1] that specially focuses on the improvement of speech generation downstream tasks. On top of the teacher-student-based training objective used in data2vec, we introduce an additional reconstruction module and a reconstruction loss that enforces the learned representations to contain enough information to reconstruct the clean speech waveforms given noisy masked input. The generative performance of our proposed SSL framework is tested for three downstream tasks; speech enhancement, speech separation and packet loss concealment. Experimental results on all three tasks show improved performance over the baselines. The major contribution can be summarized as two-fold:

1. We develop data2vec-SG that shows better performance on speech generation downstream tasks, which has been tested for speech enhancement, speech separation and packet loss concealment.
2. The proposed SSL representation is still capable of helping other downstream tasks. Experiments on ASR show that the learned representations show significant WER improvement

*Work done during an internship at Microsoft.

in noisy conditions while maintaining the performance in clean conditions.

2. DATA2VEC-SG

Before explaining our proposed framework, we briefly introduce data2vec [1], which is used as a basis of our proposed framework. In data2vec, a teacher network and a student network that have exactly the same model architecture are co-optimized with unlabeled data. While the teacher network can access to the complete input, the student network can access to only the partially masked input. During the training, the student network is optimized to predict the teacher's representations of the full input data given the masked input data. The teacher's parameters are updated by computing the exponential average of the student's parameters.

Our proposed framework, named data2vec-SG, is developed based on the teacher-student architecture of data2vec. The overview of data2vec-SG is illustrated in Fig. 1. There are two major differences from the data2vec. Firstly, we conduct online noise mixing for the speech input. More specifically, we feed noisy speech signals to the teacher network while feeding masked noisy speech to the student network. Secondly, we add a decoder on top of both networks, where the decoder performs clean waveform reconstruction based on the contextualized representations[1]. The decoder is used only for pre-training, and it is discarded in the fine-tuning stage. The parameter of the decoder is shared between teacher and student networks, which we found slightly better compared to the model using different parameters for teacher's and student's decoders. In our implementation, the decoder consists of a three-layer transformer followed by a convolutional upsampler. The upsampler is symmetric to the convolutional layer at the bottom of the shared encoder such that the reconstructed output has the same length as the original input waveform. The contextualized representations obtained from the teacher and student networks are separately fed to the decoder.

During the training, similar to data2vec, the parameters of the teacher network are updated in each training step by calculating the exponential moving average of the parameters of the student network,

$$\theta_t \leftarrow \gamma\theta_t + (1-\gamma)\theta_s, \tag{1}$$

where θ_t and θ_s are the trainable parameters in the teacher and student networks, respectively. The weight γ is linearly increased for each update until it reaches the pre-determined maximum value. This scheduling strategy makes sure the teacher gets updated more frequently at the beginning of the training, and less frequently after a while when the training is more stable and the parameters are less randomized. For updating the student network, we adopt the Smooth L1 loss proposed for data2vec,

$$\mathcal{L}^{l1} = \begin{cases} \frac{1}{2}(z(\tau) - f(\tau))^2/\beta & \text{if} |z(\tau) - f(\tau)| \leq \beta \\ |z(\tau) - f(\tau))| - \frac{1}{2}\beta & \text{otherwise} \end{cases} \tag{2}$$

where $z(\tau)$ is the average of the normalized output of top K layers of the teacher network at time step τ, and $f(\tau)$ is the corresponding model prediction obtained from the student network. The parameter β is set to 1.0 such that the loss is less sensitive to outliers. In addition to $\mathcal{L}^{l1}$, we add a reconstruction loss $\mathcal{L}^{rec}$ defined as follows.

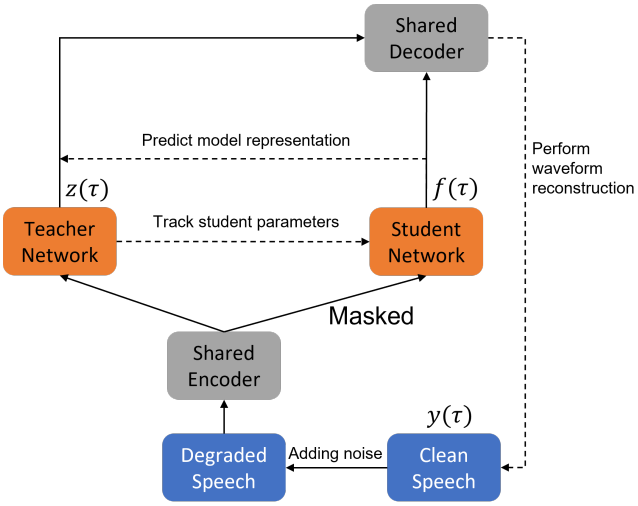

Fig. 1: The overview of data2vec-SG.

$$\mathcal{L}^{rec} = \frac{1}{T}\sum_{\tau=1}^{T}(|\hat{y}_t(\tau) - y(\tau)| + |\hat{y}_s(\tau) - y(\tau)|), \tag{3}$$

where L1 loss is computed for the reconstructed waveform obtained from the teacher network $\hat{y}_t$, the reconstructed waveform obtained from the student network $\hat{y}_s$, and the clean waveform y. The total number of time steps is denoted as T. The complete training objective of data2vec-SG is calculated as $\mathcal{L} = \mathcal{L}^{l1} + \lambda\mathcal{L}^{rec}$. For the stability of the training, we set $\lambda = 0.1$ for the first 50k iterations of the training, and then change it to 0.01.

3. EXPERIMENTS

3.1. Pre-training configuration

Our implementation is based on the official release of data2vec from FAIRSEQ [15]. We follow the Base configurations where the 960 hours of Librispeech [16] dataset is used to train the model. Our model has 95M trainable parameters for the encoder part, which is consistent with the data2vec Base architecture. The shared decoder has 23M parameters, containing 3 transformer blocks with a hidden dimension of 768, and only 1 attention head to reduce memory usage. The model is optimized with the Adam optimizer with a peak learning rate of 5e-4. A tri-stage scheduler is employed such that the learning rate is warmed up for the first 3% iterations, held for 90% of the updates and linearly decayed for the remaining part. For the noise mixing, noise files are randomly sampled from the noise set used in the deep noise suppression (DNS) challenge [17]. For each utterance, we mix the clean utterance with a noise file that is processed to be of the same length at an signal-to-noise ration (SNR) that is uniformly sampled from the range of [5, 20] dB.

3.2. Fine-tuning for generative tasks

For the fine-tuning stage, we discard the decoder module, and only use the student network for the feature extractor. Depending on the specific task, we either use the output of the last layer or the weighted average of all layers. In terms of the downstream tasks, we mainly focus on three generative tasks, speech enhancement, speech separation, and packet loss concealment.

[1]Contextualized representations here refer to the normalized top-K layers of the teacher network. This term also refers to the prediction of the student network (the last layer output), as it learns to predict contextualized representations

Table 1: SE performance on the SUPERB-SG benchmark.

Model	STOI (↑)	PESQ (↑)
No SE baseline	0.9120	1.971
FBANK	0.9364	2.553
wav2vec2 Base	0.9383	2.556
Hubert Base	0.9390	2.576
WavLM Base	0.9395	2.578
data2vec Base	0.9388	2.571
data2vec-SG Base	**0.9401**	**2.587**

Table 2: Ablation study of the proposed framework on the SE task of the SUPERB-SG benchmark.

Model	STOI (↑)	PESQ (↑)
data2vec-SG Base	**0.9401**	**2.587**
- (decoder parameter sharing)	0.9395	**2.587**
- (teacher reconstruction loss)	0.9393	2.584
- (adding noisy input for teacher)	0.9359	2.579

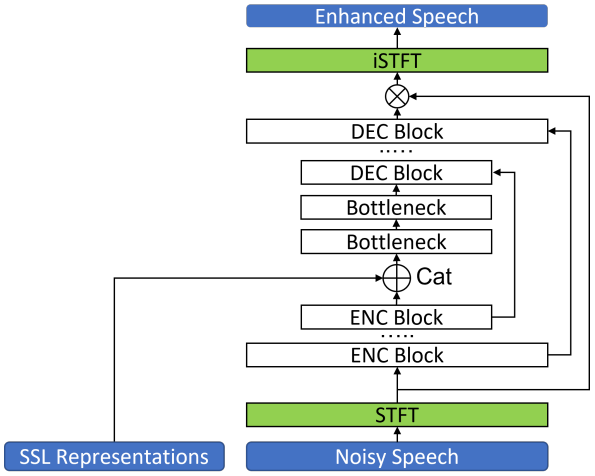

Fig. 2: The diagram illustrates the fine-tuning pipeline of SE task on the DNS dataset.

3.2.1. Speech enhancement

Speech enhancement (SE) aims to suppress the background noise from noisy speech. We have two setups for this task. The first setup follows the SE tasks in the SUPERB-SG benchmark [8], where we train and evaluate using the Voicebank-DEMAND [18] corpus. It is a benchmark dataset for comparing enhancement models, and contains 8.8, 0.6 and 0.6 hours of speech utterances for training, validation and testing, respectively. The fine-tuning of SE is conducted by employing a simple three-layer bidirectional long-short term memory (BLSTM) network that performs mask-based speech enhancement [19]. The SE performance is evaluated using STOI and PESQ. For both metrics, a higher value indicates a better SE performance.

Table 1 displays the results on the SE tasks of the SUPERB-SG benchmark. We use the last layer's output of the student network for the input feature of BLSTM-based SS model. For the fairness of comparison, all SSL-based models have roughly the same number of model parameters and pre-trained with the same 960 hours of Librispeech, as known as the Base configuration. As illustrated in the table, data2vec-SG shows a noticeable improvement in both STOI and PESQ compared to the original data2vec, where STOI is improved by 0.13% and PESQ is improved by 0.016, respectively. As a result, data2vec-SG achieves considerable improvement over the FBANK features, producing better results than the state-of-the-art SSL model WavLM.

We also examine several variants of data2vec-SG as an ablation study, whose results are presented in Table 2. The results on the 2nd row are obtained by using separate decoder parameters for the teacher and student instead of sharing the parameters of teacher's and students' decoders. The results on the 3rd row are obtained when we stop using the reconstruction loss over the teacher network. Finally, the results on the 4th row are obtained when we feed different inputs for teacher and student networks. Specifically, we feed the teacher network less noisy input by mixing the speech and noise with a higher SNR than that is used for the input of the student network within the range of [5, 20] dB. From these results, we confirm the effectiveness of the proposed model and data configuration for both STOI and PESQ.

Based on the promising results from the SUPERB-SG tasks, we then train and evaluate the proposed data2vec-SG with the second setup, which follows conventional SE configurations. We adopt the speech data and noises from the DNS challenge [20], where much larger fine-tuning data with a more sophisticated SE model architecture are utilized. Here, we use a relatively large dataset `read_speech` subset extracted from the DNS challenge for fine-tuning, which consists of around 550 hours of clean speech. In this experiment, we use the deep complex convolutional recurrent network (DCCRN) [21], which is a widely used lightweight but effective complex domain SE network. Unlike the SUPERB-SG setup that directly employs the SSL features to perform the SE task, we feed the learned representation to the bottleneck part of DCCRN. As illustrated in Fig. 2, the feature maps are concatenated with the SSL representations and then fed to the bottleneck layers. We employ the weighted sum of outputs from transformer layers of the student network as the SSL representation.

Table 3 shows the SE results on the DNS challenge 3 synthetic testing utterances (with no reverberation) [17]. For a more detailed analysis, we mix the clean utterances with two types of noises, factory noise (stationary) and babble noise (non-stationary) from the NOISEX-92 dataset [22] at two SNR levels 0 dB and 5 dB. Our results suggest that the proposed model outperforms WavLM Base and data2vec Base for both noise types at various SNR levels. The advantage is more obvious at the lower SNR. Specifically, compared with the original data2vec model, at 0 dB SNR factory noise, STOI is improved by around 1.26%, and PESQ by 0.046.

3.2.2. Speech separation

Speech separation (SS) is a task to convert multi-speaker-mixed speech signal into multiple overlap-free speech signals. In this study, we follow the two-speaker SS task in the SUPERB-SG benchmark. The experiment is conducted on the 16 kHz version of the Libri2Mix dataset, which is simulated by mixing the utterances of the Librispeech [16]. We focus on the `mix_clean` condition and each utterance contains 2 speakers. A 3-layer BLSTM model with a dimension of 896 for each direction is employed to perform the masking-based SS. We feed the last layer's output of the SSL models as an input feature to the BLSTM-based SS model. Permutation invariant training (PIT) [23] is adopted to calculate the mean-squared loss determined by the difference between the predicted mask and the ideal Non-negative Phase Sensitive Mask [19]. The separation performance is assessed by the scale-invariant signal-to-distortion ratio (SI-SDR) [24], and we display the results using the SI-SDR improvement (SI-SDRi) over the non-processed speech mixtures.

The result is presented in Table 4. With obvious improvement over the data2vec Base, we also find it outperforms WavLM. The SI-

Table 3: SE performance on the synthetic testing utterances of the DNS challenge dataset.

	Factory noise				Babble noise			
	0 dB		5 dB		0 dB		5 dB	
Model	STOI (↑)	PESQ (↑)	STOI (↑)	PESQ (↑)	STOI (↑)	PESQ (↑)	STOI (↑)	PESQ (↑)
No SE baseline	0.7033	1.0542	0.8222	1.1090	0.7085	1.0869	0.8299	1.1802
WavLM Base	0.8386	1.3604	0.9159	1.7802	0.8383	1.3734	0.9234	1.8688
data2vec Base	0.8375	1.3581	0.9153	1.7690	0.8356	1.3622	0.9214	1.8462
data2vec-SG Base	**0.8501**	**1.4037**	**0.9216**	**1.8417**	**0.8504**	**1.4149**	**0.9285**	**1.9185**

Table 4: SS performance on the SUPERB-SG benchmark.

Model	SI-SDRi (↑)
No SS baseline	-
FBANK	9.23
WavLM Base	10.37
data2vec Base	9.76
data2vec-SG Base	**10.80**

Table 5: PLC performance on the PLC challenge test dataset.

Model	PLC-MOS (↑)
Zero-filling baseline	2.90
FBANK	3.78
WavLM Base	3.83
data2vec Base	3.81
data2vec-SG Base	**3.88**

SDRi is 1.04 dB over the original data2vec, and 0.43 dB compared with WavLM Base. It is worth noting that WavLM Base incorporates speaker overlap in 20% of the training utterances during the pre-training stage, while we did not. We also found that if we perform speaker overlap, the SI-SDRi improves by 0.47 dB, but with a sacrifice in the speech enhancement performance.

3.2.3. Packet loss concealment

During signal transmission, it is unavoidable some packets are lost or arrive too late, and packet loss concealment (PLC) is employed to restore the lost packets and enables a more robust transmission system. For this downstream task, we adopt the experimental setup of the INTERSPEECH 2022 Audio Deep Packet Loss Concealment challenge[2] to evaluate the PLC performance. For fine-tuning, we conduct experiments on the dataset provided in the challenge, which contains 23184, 966 and 966 utterances for training, validation, and testing, respectively. The dataset is constructed using actual packet loss traces collected in Microsoft Teams to randomly chosen segments of audio from a podcast dataset [25]. We adopt the E3Net [26] as the backbone to perform PLC, and the pipeline of incorporating SSL representations is similar to the settings depicted in Fig. 2. Specifically, we insert SSL representations before the LSTM bottlenecks and concatenate them with the feature maps obtained from convolutional encoder. The final performance is evaluated using the PLC-MOS[3], which is the neural network-based estimator of human ratings.

We again observe the performance advantage of data2vec-SG, which is illustrated in Table 5. As shown in the table, incorporating SSL baselines all show clear improvement over the FBANK features. We observe a similar trend as we observe in the SE and SS tasks, where data2vec-SG shows a clear advantage over WavLM Base and data2vec Base. Specifically, the PLC-MOS is improved over 0.11 compared with WavLM.

[2]https://github.com/microsoft/PLC-Challenge

[3]https://github.com/microsoft/PLC-Challenge/tree/main/PLCMOS

Table 6: ASR performance on both clean and noisy LibriSpeech.

	Clean WER (↓)		Noisy WER (↓)	
Model	test-clean	test-other	test-clean	test-other
data2vec Base	**2.8**	**6.8**	37.9	55.3
data2vec-SG Base	3.0	7.0	**15.5**	**30.0**

3.3. Fine-tuning for ASR task

We also conduct clean and noisy ASR evaluations to demonstrate the generalization capability of the proposed representation for non-speech-generation downstream tasks. The fine-tuning regime follows the setup of wav2vec2.0, where a linear adaption layer is added over the learned representations to produce character predictions. We use the 100-hour subset of the Libri-light corpus [27] as the labeled data to fine-tune the model with connectionist temporal classification (CTC) loss. During the evaluation, a 4-gram language model [28] trained on LibriSpeech is applied with a beam size of 1500. The word error rate (WER) results are reported for both clean and noisy conditions using two subsets (`test-clean` and `test-other`) from Librispeech. The noisy speech for evaluation is simulated by mixing the clean utterances with noises extracted from the MUSAN corpus [29] at an SNR randomly sampled from [5, 20] dB.

As shown in Table 6, compared with the original data2vec, the WER performance is similar on the clean conditions, and we observe 0.2% absolute WER degradation on the `test-clean` subset. This result demonstrates the effectiveness of our approach to improving noise robustness of the learned representation without sacrificing the capabilities of clean speech.

4. CONCLUSION

In this paper, we proposed data2vec-SG which specially focuses on the improvement of speech generation downstream tasks. On top of the teacher-student learning framework proposed in data2vec, we added a reconstruction module that enforces the representations to contain enough information to generate clean speech waveforms. Experimental results showed that our proposed model achieved better performance on generative tasks like speech enhancement, speech separation, and packet loss concealment. Meanwhile, the representation still generalized well to other tasks. For instance, it benefited ASR and shows good WER results in both clean and noisy conditions. For future work, we plan to leverage both labeled and unlabeled noisy speech during the pre-training stage, such that we can relax the constraint that the clean speech label is required.

5. REFERENCES

[1] 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:2202.03555*, 2022.

[2] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *Proceedings of INTERSPEECH*, pp. 146–150, 2019.

[3] A. T. Liu, S.-W. Yang, P.-H. Chi, P.-C. Hsu, and H.-Y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proceedings of ICASSP*, 2020, pp. 6419–6423.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

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

[6] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, et al., "Unispeech-SAT: Universal speech representation learning with speaker aware pre-training," in *Proceedings of ICASSP*, 2022, pp. 6152–6156.

[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[8] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-W. Yang, S. Dong, A. T. Liu, C.-I.J Lai, J. Shi, et al., "SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities," in *Proceedings of ACL*, 2022, pp. 8479–8492.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. on ASLP*, vol. 19, pp. 2125–2136, 2011.

[10] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of ICASSP*, 2001, pp. 749–752.

[11] Zili Huang, Shinji Watanabe, Shu-wen Yang, Paola García, and Sanjeev Khudanpur, "Investigating self-supervised learning for speech enhancement and separation," in *Proceedings of ICASSP*, 2022, pp. 6837–6841.

[12] K.-H. Hung, S.-W. Fu, H.-H. Tseng, H.-T. Chiang, Y. Tsao, and C.-W. Lin, "Boosting self-supervised embeddings for speech enhancement," *arXiv:2204.03339*, 2022.

[13] Saurabh Kataria, Jesús Villalba, and Najim Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," in *Proceedings of ICASSP*, 2021, pp. 7118–7122.

[14] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *Proceedings of ICASSP*, 2020, pp. 6989–6993.

[15] Myle O., Sergey E., Alexei B., Angela F., Sam G., Nathan N., David G., and Michael A., "FAIRSEQ: A fast, extensible toolkit for sequence modeling," in *Proceedings NAACL-HLT: Demonstrations*, 2019.

[16] V. Panayotov, G. Chen, D. Povey, and S. Hudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of ICASSP*, 2015, pp. 5206–5210.

[17] C.KA Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proceedings of INTERSPEECH*, 2020, pp. 2492–2496.

[18] Christophe Veaux, Junichi Yamagishi, and Simon King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proceedings of O-COCOSDA/CASLRE*. IEEE, 2013, pp. 1–4.

[19] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. on ASLP*, vol. 25, no. 10, pp. 1901–1913, 2017.

[20] C.KA Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *Proceedings of ICASSP*, 2021, pp. 6623–6627.

[21] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proceedings of INTERSPEECH*, 2020, pp. 2482–2486.

[22] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[23] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proceedings of ICASSP*. IEEE, 2017, pp. 241–245.

[24] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proceedings of ICASSP*, 2019, pp. 626–630.

[25] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "INTERSPEECH 2022 audio deep packet loss concealment challenge," *arXiv:2204.05222*, 2022.

[26] M. Thakker, S.E. Eskimez, T. Yoshioka, and H. Wang, "Fast real-time personalized speech enhancement: End-to-end enhancement network (E3Net) and knowledge distillation," *arXiv:2204.00771*, 2022.

[27] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, et al., "Libri-light: A benchmark for ASR with limited or no supervision," in *Proceedings of ICASSP*, 2020, pp. 7669–7673.

[28] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the statistical machine translation*, 2011, pp. 187–197.

[29] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv:1510.08484*, 2015.