INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Using Semi-supervised Learning for Monaural Time-domain Speech Separation with a Self-supervised Learning-based SI-SNR Estimator

Shaoxiang Dang[1], *Tetsuya Matsumoto*[2], *Yoshinori Takeuchi*[3], *Hiroaki Kudo*[4]

[124]Graduate School of Informatics, Nagoya University, Japan
[3]School of Informatics, Daido University, Japan
dang.shaoxiang.s0@s.mail.nagoya-u.ac.jp

Abstract

Speech separation aims to decompose mixed speeches into independent signals. Prior research on monaural time-domain speech separation has made great progress in supervised manners. Almost all of these works are trained on simulated mixed speech signals since obtaining ground truth for real-world mixed signals is problematic. To this end, we propose a novel semi-supervised learning method for speech separation (SSLM-SS), which leverages mixed speeches without ground truth. In particular, for this type of data, we further put forward a non-intrusive separated speech quality prediction network (SSQP-Net) based on self-supervised learning. According to the results, the linear correlation coefficient between the predicted results of SSQP-Net and the ground truth achieves 0.9. Moreover, the performance of SSLM-SS equipped with SSQP-Net exhibits an improvement of 0.2 dB and 1.1 dB compared to the mixture invariant training (MixIT) in the conditions of involving 10% and 50% labeled data respectively, and rivals fully supervised learning.

Index Terms: monaural speech separation, semi-supervised learning, self-supervised learning, non-intrusive SI-SNR estimation

1. Introduction

Speech separation has emerged as a highly sought-after topic in speech-related fields. In recent years, numerous speech separation models (SSMs) have demonstrated impressive performance on the scale-invariant signal-to-noise ratio (SI-SNR) using supervised learning approaches [1, 2]. Notable models include Conv-Tasnet (CTN) [3], Dual-path Recurrent Neural Network (DR) [4], and SepFormer (SF) [5]. These models have been trained and evaluated on public data sets such as WSJ0-2Mix [6] and LibriMix [7], where mixed signals are synthesized using speech samples from the Wall Street Journal (WSJ) and LibriSpeech [8]. However, synthetic data sets suffer from a common limitation: they lack realism. To address this issue, the WHAM! [9] and WHAMR! [10] datasets have been introduced, which incorporate real-world noise and reverberation into synthesized signals. Despite the significant progress made, there still remains a gap between synthetically mixed speeches and real-world mixed speeches [11, 12]. Mixture invariant training (MixIT) has been proposed for training speech separation models without relying on ground truth. During training, MixIT remixes the inputs of mixed signals, creating a Mixture of Mixture (MoM), and uses the mixed signals as targets to separate MoM [8]. However, the high accuracy achieved by MixIT is attributed to the inclusion of a large amount of non-mixed data in the training dataset, which leads to a training approach that resembles common supervised methods.

On the other hand, the remarkable results from prior studies on the objective assessment of the perceptual evaluation of speech quality (PESQ) [13] and the mean opinion score (MOS), the commonly used metrics in speech enhancement and voice conversion, demonstrate the potential for high-quality and non-intrusive objective speech assessment in a wide range of settings [14, 15, 16, 17]. One approach utilizing convolutional neural networks (CNNs) to predict the SI-SNR value from the mixture and separated signals performs well [11]. However, it is considered insufficient for two reasons. Firstly, the target SI-SNR value used to train the prediction model was compressed between 0-10 dB, suggesting that the reported error could potentially be much greater at the actual scale. Secondly, the model assesses one separated signal from the mixed signal at a time. While this approach maintains consistency with the original SI-SNR calculation formula, processing all separated speech signals simultaneously could provide convenience in the context of this paper where it is required to predict SI-SNRs for multiple separated speech signals.

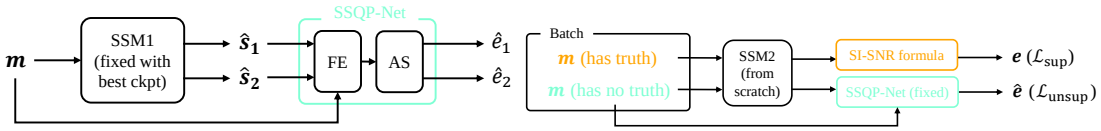
Building upon the Asteroid framework [18], we introduce a novel semi-supervised learning method for monaural time-domain speech separation (SSLM-SS) embracing a pre-trained SI-SNR estimator which is referred to as non-intrusive separated speech quality prediction network (SSQP-Net) and provides supervision for unlabeled signals. Instead of learning speech features from scratch, SSQP-Net utilizes a pre-trained module to extract robust speech features from a large-scale dataset of unlabeled speech signals via self-supervised representation learning (SSL) [19]. After coupling this with an SI-SNR prediction module, the entire network is fine-tuned on the SI-SNR estimation task. According to our experimental results, SSQP-Net can outperform previous work [11]. Most importantly, different from this prior work that compresses SI-SNR value in 0-10 dB, SSQP-Net can predict the separated speech at true scale, for example, separated signals with greater SI-SNR values than 10 dB, which is pivotal to provide reliable labels during semi-supervised learning for signals without ground truth. By using SSQP-Net, SSLM-SS can achieve better performance than the fully supervised method. Moreover, we compare SSLM-SS with MixIT. Experimental results prove the superiority of SSLM-SS in the same condition. We eventually validate our work on the real-life dataset LibriCSS [20], and the results consistently prove the effectiveness of the proposed method.

2. Related knowledge

2.1. SI-SNR

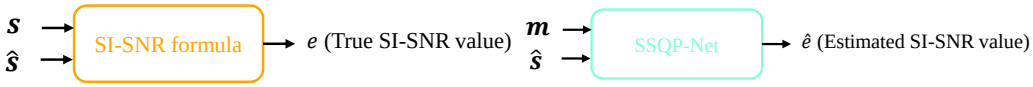
In previous monaural time-domain speech separation models based on supervised learning, SI-SNR e is commonly used to

10.21437/Interspeech.2023-85



(a) *SSQP-Net training flow chart. We use a fixed SSM with the best checkpoint to train the SSQP-Net.* $\boldsymbol{m}$ *is the mixed signal.*

(b) *SSLM-SS on batch level. We use the pre-trained SSQP-Net to provide estimated SI-SNR values and train the SSM2.*



(c) *Supervised part.*

(d) *Unsupervised part.*

Figure 1: *Overall structure of SSLM-SS. In our work, the SSM1 with the best checkpoint used in (a) is DR, and the SSM2 to be trained from scratch in (b) is CTN. Tangerine and aquamarine depict the supervised and unsupervised parts respectively in the entire SSLM-SS structure.*

measure the similarity of separated signal $\hat{\boldsymbol{s}}$ and corresponding clean reference $\boldsymbol{s}$ as follows [2]:

$$e = 10\log_{10}\frac{\| \boldsymbol{s_p} \|^2}{\| \hat{\boldsymbol{s}} - \boldsymbol{s_p} \|^2} \tag{1}$$

where $\boldsymbol{s_p}$ denotes the projection of $\hat{\boldsymbol{s}}$ on $\boldsymbol{s}$:

$$\boldsymbol{s_p} = \frac{\langle \hat{\boldsymbol{s}}, \boldsymbol{s} \rangle \boldsymbol{s}}{\| \boldsymbol{s} \|^2} \tag{2}$$

2.2. SSL

Directly modeling speech for a specific task may not lead to the discovery of universal features, such as contextualized representations. However, these representations can be learned through SSL by training the model to distinguish the target sample from distracting samples [19, 21, 22, 23]. This is formulated as:

$$\boldsymbol{Z} = h_1(\boldsymbol{x}) \tag{3}$$
$$\boldsymbol{H} = h_2(m(\boldsymbol{Z})) \tag{4}$$

where $h_1(\cdot)$ is a series of CNNs that extracts information from the input waveform $\boldsymbol{x}$ in a fixed length, $h_2(\cdot)$ is a transformer that encodes the masked localized representations into contextualized representations, and $m(\cdot)$ is a masking operation. The final step in various works differs greatly. Some works use an additional quantization operation $q(\cdot)$ to generate learned representations from unmasked localized representations by computing $\boldsymbol{Q} = q(\boldsymbol{Z})$ and force the system to distinguish $\boldsymbol{q_t}$ from a set of elements of $\boldsymbol{Q}$ given $\boldsymbol{h_t}$ [21]. Others produce temporary ground truth $\boldsymbol{C}$ by clustering MFCC features beforehand and prompt the model to discriminate $\boldsymbol{c_t}$ from a set of elements of $\boldsymbol{C}$ given $\boldsymbol{h_t}$ [22, 23].

3. Methods

3.1. SSLM-SS

In this work, we focus on 2-speaker separation. Fig. 1 illustrates the entire picture of SSLM-SS from the batch's perspective. In each batch of training, the SI-SNR formula calculates SI-SNR values for separated signals that have ground truth available, while SSQP-Net predicts SI-SNR values for separated signals that don't have ground truth. The cost function of SSLM-SS is as follows

$$\mathcal{L}_{\text{SSLM-SS}} = (1 - f(t))\mathcal{L}_{\text{sup}} + f(t)\mathcal{L}_{\text{unsup}} \tag{5}$$

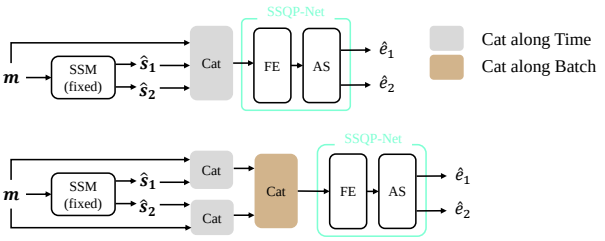


Figure 2: *Diagrams of OTFP (top) and OTFS (bottom).*

where $f(t)$ denotes the weight function, and t denotes the number of training batch.

3.2. SSQP-Net

Extending the previous SI-SNR estimator in [11] that takes one separated signal and the mixture as inputs at a time, which is denoted as one-time-for-single (OTFS), proposed SSQP-Net also checks another straightforward modeling approach one-time-for-pair (OTFP) that processes two separated signals as a pair. Notably, the first step for both OTFS and OTFP in SSQP-Net is concatenation along the time dimension, which differs from the previous one [11] that used CNNs, where concatenation is performed along the channel dimension. OTFS and OTFP are illustrated in Fig. 2. The SSQP-Net consists of a feature extraction (FE) module and an SI-SNR assessment (AS) module.

3.2.1. FE module

FE module aims to extract features from speech signals. We construct the FE module through OTFP and OTFS as follows:

$$\boldsymbol{F} = g_{\text{otfp}}(\boldsymbol{m}, \hat{\boldsymbol{s_1}}, \hat{\boldsymbol{s_2}}) \tag{6}$$
$$\boldsymbol{F} = g_{\text{otfs}}(\boldsymbol{m}, \hat{\boldsymbol{s}}) \tag{7}$$

where $g_{\text{otfp}}(\cdot)$ and $g_{\text{otfs}}(\cdot)$ describe FE modules pre-trained via SSL and shape SSQP-Net in OTFP and OTFS manners, respectively. $\hat{\boldsymbol{s_1}}$ and $\hat{\boldsymbol{s_2}}$ represent the two separated signals from the same mixture signal $\boldsymbol{m}$, while $\hat{\boldsymbol{s}}$ represents each of the separated signals from $\boldsymbol{m}$. The extracted features are represented by $\boldsymbol{F} \in \mathbb{R}^{d_{\text{feature}} \times d_{\text{frame}}}$.

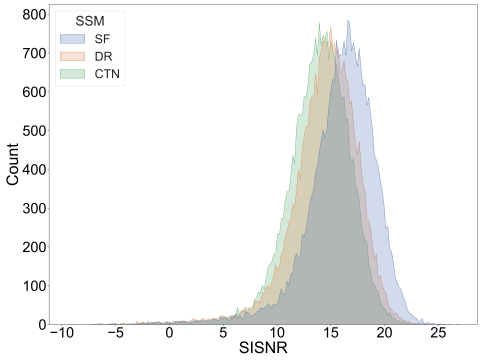


Figure 3: *The distributions of training data of the best three checkpoints using different SSMs.*

3.2.2. SI-SNR AS module

With the powerful feature extraction module obtained through SSL, the SI-SNR assessment module is constructed simply by performing an average operation over frames and applying a linear transformation:

$$\hat{\boldsymbol{e}} = \boldsymbol{W}(\frac{1}{N}\boldsymbol{F}\boldsymbol{I}) + \boldsymbol{b} \tag{8}$$

where $\boldsymbol{I} \in \mathbb{R}^{d_{\text{frame}} \times 1}$ represents the matrix of ones, N represents the number of frames. $\boldsymbol{W} \in \mathbb{R}^{d_{\text{output}} \times d_{\text{feature}}}$ and $\boldsymbol{b} \in \mathbb{R}^{d_{\text{output}} \times 1}$ are weight and bias respectively. The dimension of d_{output} is determined by the type of FE module.

4. Experiment configuration

The experiment consists of SSQP-Net and SSLM-SS parts, and both parts are trained on NVIDIA GeForce RTX 3080 Ti. All datasets are from Libri2Mix with a sampling rate of 8 kHz.

4.1. Experiment of SSQP-Net

Following the data preparation method described in [11], we adopt the best checkpoints in each of the three SSMs: SF, DR, and CTN. We use these checkpoints to separate the mixture signals of *train-100* and use clean references to calculate the true SI-SNR of every sample. These separated signals and corresponding SI-SNR values are used as the training set to train SSQP-Net. Likewise, validation and test sets are created by using three SSMs to separate subset *dev* and *test* of Libri2Mix. The distribution of each training set is shown in Fig. 3.

This study will explore Wav2vec 2.0 (wav2vec2), HuBERT, and WavLM models for FE module, which have been trained on large-scale unlabeled datasets LibriSpeech, Libri-light, LibriVox, and Mix 94k hr through SSL. Each of these models has three available scale types, and the basic one, with features in 768 dimensions, is chosen. The cost function is the mean absolute error (MAE). During fine-tuning, the warm-up learning strategy [24] is employed, where the learning rate is increased to 0.0001 in the first 10 epochs, and then halved if there is no improvement on the validation dataset for 4 consecutive epochs. In the trials with the OTFP manner, predicted SI-SNRs are averaged over sources.

For evaluation, we use linear correlation coefficient (LCC) to make a comparison with a prior model in [14], and MAE is used to compare two modeling manners: OTFP and OTFS.

4.2. Experiment of SSLM-SS

We use SSQP-Net trained on the dataset generated by the best checkpoint of DR with wav2vec2 in the OTFP manner to predict SI-SNR values. The SSLM-SS is then trained on a subset of *train-360* that has 13900 utterances and validated on *dev*. The CTN is used as the SSM in SSLM-SS. It is worth noting that CTN here will be trained from scratch. We examine three ratios using labeled signals in 10%, 50%, and 100%. We use the utterance permutation invariant training method (uPIT) to process labeled signals [25]. The number of training epochs is 150, from which we design the weight function as

$$f(t) = t/300 \tag{9}$$

For the hyper-parameters of CTN, we use the window size of 16 samples and the hopping size of 8 samples. For separation blocks of CTN, we use 6 blocks and repeat 3 times. Each sample is further segmented with 3s long before being fed to the CTN model. The learning rate is initialized at 0.001 and halved every five epochs when no dropping on the validation dataset.

We first evaluate our work on *test* set of Libri2Mix and report the results in SI-SNRi. This metric will filter out the effects of mixing conditions. Furthermore, we evaluate our work on the real-life dataset LibriCSS which derives from LibriSpeech and simulates a conversation by reading the corpus sentences alternately. It involves six subsets with varying overlapping ratios, and we exclude two subsets where overlapping ratios are 0. There are no references to clean utterances in LibriCSS, transcriptions are instead provided as labels. We use the single-channel utterance-wise evaluation schemes, except that we use a different SSL-based ASR model [21]. The recognized results will be evaluated in word error rate (WER).

5. Results and discussion

5.1. Results of SSQP-Net

5.1.1. Comparison with baseline

Table 1 shows the results in LCC. First, reported LCCs of baseline are rough 0.80, while SSQP-Net can yield better LCCs than baseline on both match and mismatch conditions. On match conditions, the trials with the training dataset generated by DP achieve the best and most stable performance of rough 0.90. All trials with FE of WavLM obtain the best LCCs on three test sets.

Table 1: *LCC results of SI-SNR predictions.*

Models	**SSM**	**FE**	**Test sets**		
			SF	DR	CTN
Baseline [11]	SF	CNNs	0.80	0.80	0.81
SSQP-Net	SF	wav2vec2	0.87	0.89	0.85
		HuBERT	0.85	0.86	0.82
		WavLM	0.87	0.90	0.86
	DR	wav2vec2	0.86	0.90	0.87
		HuBERT	0.84	0.89	0.84
		WavLM	0.86	0.90	0.87
	CTN	wav2vec2	0.85	0.89	0.88
		HuBERT	0.84	0.89	0.89
		WavLM	0.86	0.90	0.89

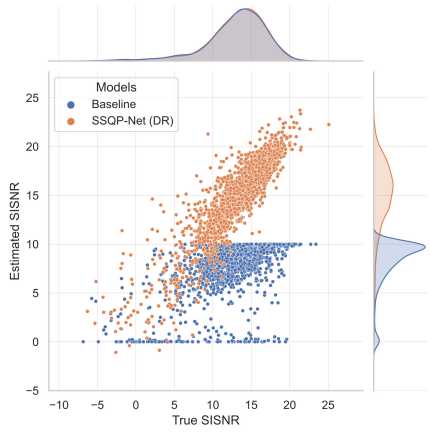


Figure 4: *The scatter plot of the baseline and SSQP-Net (DR) on test dataset generated by well-trained DR model.*

Table 2: *MAE (dB) results of wav2vec2 using OTFP and OTFS.*

Type	SSM	Test sets SF	DR	CTN
OTFP	SF	1.39	2.02	1.77
	DR	1.28	1.39	1.33
	CTN	1.59	1.28	1.20
OTFS	SF	1.30	2.00	1.57
	DR	1.36	1.31	1.29
	CTN	1.49	1.62	1.20

Since WavLM introduces a mechanism of gated relative position bias and further pre-trains on overlapping speeches, it can provide more appropriate features than wav2vec2 or HuBERT.

We exhibit a scatter plot of the baseline and SSQP-Net, the model trained on the training dataset generated by DR, on the test set generated by DR in Fig. 4, from which we can intuitively know the advantages of SSQP-Net (DR) on separated signals with high SI-SNRs, where baseline model is unable to make a precise prediction. This is also the reason why we don't directly use the baseline model to provide SI-SNR values.

5.1.2. Comparison of OTFP and OTFS

Following experiments using OTFP and FE of wav2vec2, we further examine the way of OTFS, and results on MAE are reported in Table 2. We observe that both approaches produce close outcomes. In most cases, trials with OTFS can get slightly better MAE than those with OTFP. This suggests that the combined features of all separated signals from the same mixture are not necessary to predict SI-SNR values whose original calculation formula also excludes features from the opponent.

5.2. Results of SSLM-SS

Table 3 displays the SI-SNRi results of MixIT and SSLM-SS. By using the fully supervised training approach, the CTN model achieves 13.0 dB, while this number drops drastically to 8.8 dB when the amount of training data is shrunk to one-tenth. In the trial where the proportion of labeled data is 10%, the result of MixIT is 11.5 dB. In contrast, SSLM-SS can yield a slight improvement of 0.2 dB. This figure is enlarged to 1.1 dB in trials

Table 3: *SI-SNRi (dB) results of baselines and SSLM-SS. In implementing of CTN, % supervised data indicates the proportion of data used for fully supervised training.*

Models	% supervised data 10%	50%	100%
CTN	8.8	11.6	13.0
MixIT [26]	11.5	11.8	-
SSLM-SS	**11.7**	**12.9**	-

using 50% labeled data. The performance decreases dramatically when the labeled data only accounts for 10%, a possible reason is that the proposed SSQP-Net cannot accurately process signals with small SI-SNR data, and the model is fed samples of unreliable SI-SNRs in the early stage of training. The trade-off between the accurate prediction of high and small SI-SNR data will be further investigated in future work.

Finally, SSLM-SS is examined on LibriCSS in Table 4. In subsets with high overlapping ratios, SSLM-SS reduces the WERs compared to mixture whose results are obtained by recognizing the mixed speeches. SSLM-SS (50%) and SSLM-SS (10%) outperform CTN (50%) and CTN (10%) respectively. Specifically, SSLM-SS (10%) win CTN (10%) 2.8% on the subset with a 40% overlapping ratio. Aiming at fully supervised CTN (100%), the proposed SSLM-SS (50%) reaches closer WERs on the subsets with varying overlapping ratios.

Table 4: *WER (%) results on LibriCSS.*

Models	LibriCSS (overlap ratio in%) 10	20	30	40
Mixture	9.2	15.3	24.4	32.5
CTN (100%)	9.4	12.5	18.1	21.3
CTN (50%)	11.3	14.9	20.4	25.0
SSLM-SS (50%)	**11.2**	**14.5**	**20.1**	**23.2**
CTN (10%)	12.7	16.8	24.1	28.5
SSLM-SS (10%)	**12.1**	**15.6**	**21.1**	**25.7**

6. Conclusion

In this work, we have put forward a novel semi-supervised training method for speech separation (SSLM-SS). Different from previous models, SSLM-SS contains an SSQP-Net part that can predict the SI-SNR value for separated signals in the absence of clean references. SSQP-Net outperforms the related work by about 10% on LCC. Most importantly, SSQP-Net can precisely predict separated signals with high SI-SNR, which is super desired for SSLM-SS. Equipping with SSQP-Net, SSLM-SS shows the powerful ability to achieve close performance as fully supervised training at the signal level, which is also superior to MixIT. Through a further examination on LibriCSS, SSLM-SS performs consistently great on real mixed signals. A promising future work is an investigation of training SSQP-Net on data with diverse SI-SNR values.

7. Acknowledgement

The first author of this work is supported by Nagoya University CIBoG WISE program from MEXT.

8. References

[1] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[2] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[3] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[4] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[5] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[6] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[7] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[9] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[10] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 696–700.

[11] C. Subakan, M. Ravanelli, S. Cornell, and F. Grondin, "Real-m: Towards speech separation on real mixtures," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6862–6866.

[12] A. Sivaraman, S. Wisdom, H. Erdogan, and J. R. Hershey, "Adapting speech separation to real-world meetings using mixture invariant training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 686–690.

[13] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[14] M. Yu, C. Zhang, Y. Xu, S. Zhang, and D. Yu, "Metricnet: Towards improved modeling for non-intrusive speech quality assessment," *arXiv preprint arXiv:2104.01227*, 2021.

[15] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," *arXiv preprint arXiv:1904.08352*, 2019.

[16] W.-C. Tseng, C.-y. Huang, W.-T. Kao, Y. Y. Lin, and H.-y. Lee, "Utilizing self-supervised representations for mos prediction," *arXiv preprint arXiv:2104.03017*, 2021.

[17] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8442–8446.

[18] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[19] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *arXiv preprint arXiv:2205.10643*, 2022.

[20] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7284–7288.

[21] A. Baevski, Y. 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. 12 449–12 460, 2020.

[22] 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.

[23] 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.

[24] P. Goyal, P. Dollár, R. B. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch SGD: training imagenet in 1 hour," *CoRR*, vol. abs/1706.02677, 2017. [Online]. Available: http://arxiv.org/abs/1706.02677

[25] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[26] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3846–3857, 2020.