REAL-M: TOWARDS SPEECH SEPARATION ON REAL MIXTURES

Cem Subakan[1,2]*, Mirco Ravanelli*[2]*, Samuele Cornell*[3]*, François Grondin*[1]

[1]Université de Sherbrooke, Canada, [2]Mila-Quebec AI Institute, Canada,
[3]Università Politecnica delle Marche, Italy

ABSTRACT

In recent years, deep learning based source separation has achieved impressive results. Most studies, however, still evaluate separation models on synthetic datasets, while the performance of state-of-the-art techniques on in-the-wild speech data remains an open question. This paper contributes to fill this gap in two ways. First, we release the REAL-M dataset, a crowd-sourced corpus of real-life mixtures. Secondly, we address the problem of performance evaluation of real-life mixtures, where the ground truth is not available. We bypass this issue by carefully designing a blind Scale-Invariant Signal-to-Noise Ratio (SI-SNR) neural estimator. Through a user study, we show that our estimator reliably evaluates the separation performance on real mixtures, i.e. we observe that the performance predictions of the SI-SNR estimator correlate well with human opinions. Moreover, when evaluating popular speech separation models, we observe that the performance trends predicted by our estimator on the REAL-M dataset closely follow the performance trends achieved on synthetic benchmarks.

Index Terms— Source separation, In-the-wild speech separation, Dataset, Blind SI-SNR estimation, Deep learning.

1. INTRODUCTION

Source separation techniques have evolved quickly in the last few years and recently reached impressive performance levels. SepFormer [1], Wavesplit [2], and DualPath RNN [3] (DPRNN), for instance, can achieve more than 20 dB improvement in the Scale Invariant Signal-to-Noise Ratio (SI-SNR) [4]. The vast majority of studies in the field employ synthetic datasets for both training and evaluation purposes [5–8]. This practice is largely accepted by the community and offers undeniable advantages, as simulated data are easily created from clean recordings. The widely used WSJ0-2/3Mix datasets [9], which are generally considered the de-facto standard benchmark, consist of signals that are artificially produced by mixing speech recordings from the Wall Street Journal (WSJ) corpus [10]. The original utterances are recorded with high-quality microphones in controlled environments where there is almost no noise and reverberation. Another noteworthy advantage is that artificial mixtures can be synthesized on the fly using dynamic mixing [2, 11]. This feature is extremely valuable when training neural models, as we can augment the dataset and improve the separation performance significantly.

Simulated mixtures are commonly used for evaluation purposes as well. Despite being widely adopted, we believe that this practice is potentially misleading. It might bias the community towards systems that work well in laboratory conditions and fail in real-life scenarios. Some efforts have been recently devoted to mitigating this lack of realism. For instance, WHAM! [12] and WHAMR! [13] datasets introduced environmental noise and reverberation, but still rely on simulated data for evaluation that are not entirely representative of the challenges faced with in-the-wild signals. Other examples in this direction are real-world datasets based on multi-party meeting scenarios such as [14–17]. However these datasets mainly consist of non-overlapped speech.

Evaluation through real-life mixtures is arguably the most accurate approach to avoid mismatches between laboratory and real-life conditions. The main challenge that prevents it is the lack of ground truth signals (i.e., the clean sources used as targets for training and evaluating speech separation). Specialized hardware and controlled recording conditions can circumvent this issue, but this practice is costly, time-consuming, and impractical to adopt widely. As an alternative solution in [18], authors explore the idea of using downstream tasks such as speaker verification for evaluation.

In this paper, we address the problem from a different angle and provide a two-fold contribution. First, we release REAL-M, a real-life speech source separation dataset for two-speaker mixtures with large amount of overlapping speech. We collected this dataset through crowdsourcing by asking contributors to read a predefined set of sentences simultaneously. The mixtures are recorded in different acoustic environments using a wide variety of recording devices such as laptops and smartphones, thus reflecting more closely potential application scenarios. REAL-M currently contains more than 1400 speech mixtures from 50 native and non-native English speakers, for a total of almost three hours of real-world speech mixtures with their corresponding transcriptions.

As a second contribution, we carefully design and study a neural network that blindly estimates the quality of a signal processed by a speech separation model. The idea of automatically assessing speech quality using a neural network has been recently explored in the literature. In [19–23], for instance, authors designed neural networks that estimate non-differentiable speech quality metrics (e.g., PESQ) for improving speech enhancement. These works showcase that estimating speech quality without having access to the ground truth is feasible. We here inherit the same philosophy in the context of speech separation. In particular, we employ a convolutional model to predict the SI-SNR from the audio mixture and separated signals. We explore the 2-speaker case, but the proposed framework can be easily extended to handle more speakers.

To the best of our knowledge, this paper is the first work showing that neural metric learning is reliable enough for accurately assessing speech separation on real mixtures. We show that the output of the employed SI-SNR estimator correlates well on the REAL-M dataset with human opinion scores. We also evaluated popular speech separation models using both the standard simulated data and REAL-M. The predictions of the SI-SNR estimator on REAL-M follow a trend that closely resembles the performance observed on simulated data, giving further confirmation on the effectiveness of the proposed approach.

We also release the training script of the SI-SNR estimator

within the SpeechBrain [24], and the pre-trained SI-SNR estimator on Huggingface. The dataset is available from our website.

2. DATA COLLECTION

The data collection has been conducted as follows. We showed the participants the text of two sentences randomly sampled from the test set of LibriSpeech [25] dataset. We limited the length of the sentences to be between 5 to 15 words. In total, we presented 569 unique pairs of sentences to the participants, and we collected 1436 mixtures. We chose sentence pairs with closest number of words to maximize speech overlap between speakers. We asked the contributors to simultaneously read the shown utterances while being physically in the same room. Beyond that, we also recorded 144 mixtures where one participant was recorded through a videoconferencing software (e.g., Zoom, google meet). This increases the diversity of the dataset and addresses an application scenario that frequently occurs.

Thanks to crowd-sourcing, the acoustic conditions and recording equipment encompass a wide variety of scenarios: mixtures contain varying levels of reverberation and noise as the speech can be either near or far-field. Moreover, we involved both native and non-native speakers with different accents, including American, British, French, Italian, Persian, Indian, and African. The recordings have been performed with a purposely built data collection platform interfaced with Amazon Mechanical Turk.

3. BLIND NEURAL SI-SNR ESTIMATION

As humans, we can tell if the quality of a speech signal is good or not just by listening to it. By following this reasoning, we describe here how we designed a neural network that estimates the separation performance in terms of SI-SNR without accessing the ground truth signals.

3.1. Training

The basic components needed to train the neural SI-SNR estimator are a pretrained separation model, a synthetic dataset, and a neural network for SI-SNR estimation. The training pipeline, shown in Fig. 1, comprises of the following steps:

1. First, we process the synthetic mixtures x with a pretrained speech separation model, as reported in the following equation:
$$\hat{s}_1, \hat{s}_2 \;=\; \text{PT-S}(x). \tag{1}$$
This step provides separated signals $\hat{s}_1$ and $\hat{s}_2$ (with different levels of distortions) that the neural estimator will assess. The parameters of the pretrained separator are kept frozen.
2. Then, we use the ground truth sources s_k and the separated sources $\hat{s}_k$ to compute the oracle SI-SNR values:
$$SNR_k \;=\; \text{SI-SNR}(s_k, \hat{s}_k),\; k \in \{1, 2\}. \tag{2}$$
As shown in Figure 1, the permutation over ground truth sources is resolved before calculating the oracle SI-SNR values. The oracle SI-SNR values SNR_k are used as a target for the SI-SNR estimator.
3. We feed the separated signals $\hat{s}_k$ into the neural estimator, which aims to predict the SI-SNR performance:
$$\widehat{SNR}_k \;=\; \text{SI-SNR-Estimator}(x, \hat{s}_k),\; k \in \{1, 2\} \tag{3}$$
The model is fed by the mixture signal as well. This addition leads to more accurate predictions, as it provides a more accurate guideline for the neural estimator.
4. The neural estimator is trained to regress the oracle SI-SNR values, as shown in the following equation:
$$\mathcal{L} \;=\; \|SNR_1 - \widehat{SNR}_1\|_1 + \|SNR_2 - \widehat{SNR}_2\|_1, \tag{4}$$
where $SNR_{1,2}$ are the oracle SI-SNRs computed in step 2 and $\widehat{SNR}_{1,2}$ are the estimated SI-SNR values computed at step 3. Training is conducted in a standard way using back-propagation coupled with the Adam optimizer. More details on the training pipeline can be found on the SI-SNR estimator recipe released in SpeechBrain.

3.1.1. Synthetic Mixture Creation

The training pipeline involves a synthetic dataset composed of artificial mixtures with their corresponding ground truths. In this work, we consider the LibriMix [26], and the WHAMR! [13] datasets simultaneously by randomly choosing mixtures from the two datasets.

More precisely, we create the mixtures on the fly using dynamic mixing [2, 11]. We randomly sample relative mixing SNRs in the range between 0-5dB. Environmental noise (using noisy sequences from the WHAM! corpus [12]) and reverberation (using the impulse responses of the WHAMR! [13] dataset) are added as well. The synthetic mixtures are processed by the pretrained separation models described in the following sub-section.

3.1.2. Pretrained Separation Models

We consider two different ways of using pretrained separators in the training pipeline:

- (*single*). We train the SI-SNR estimator using only one separation model. In our case, we used the SepFormer model pretrained on the WHAMR! dataset from the SpeechBrain Hugging Face repository [1].
- (*pool*). We train the SI-SNR estimator using a pool of many source separation models. Namely, we used a mixture of SepFormer [1], DPRNN [3] and Convtasnet [5] models. At training time, one model from the pool is uniformly sampled and applied to the mixture to derive the estimated sources. For each model, we use the checkpoints after the first and last epochs. Moreover, we use a checkpoint in the middle of training. In total, we use a combination of 9 pretrained separators. The *pool* approach leads to wide variability in the SI-SNR values. The estimator can observe different speech distortions and artifacts, with benefits on its generalization properties.

3.2. Inference

At inference time, we can simply feed the separated signal and the corresponding mixture to the neural estimator. The latter will provide an estimate of the SI-SNR performance. The proposed approach is blind because it does not require accessing the clean ground truth sources at inference time. It also has the advantage of being very straightforward and light-weight.

3.3. Architecture

The SI-SNR estimator consists of five convolutional layers (with a kernel size of 4, 128 channels, and stride of 1), followed by a statistical pooling layer, and two fully connected layers (with 256 neurons).

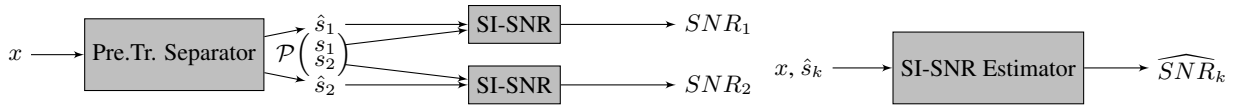

Fig. 1. Training the neural SI-SNR estimator. A pretrained separator is used to estimate the sources (left). The SI-SNR is computed using the ground truth signals (center). The SI-SNR Estimator is fed by the mixtures and the estimated sources and predicts the SI-SNRs (right).

We use ReLU activations for all the layers. We limit the estimated SI-SNR values to fall between 0 and 10 dB. This range covers the typical SI-SNRs observed on reverberant source separation datasets such as WHAMR! [13]. We also normalize the network output to fall between 0-1, where 0 is associated with 0 dB and 1 is associated with 10 dB. This range compression is performed with a sigmoid applied at the output of the network. The source estimates and the mixture are normalized to have zero-mean and unit-variance before inputting them to the SI-SNR estimator. These normalization steps are important to ensure an accurate prediction and a fast convergence of the model. The estimator employs about 300K trainable parameters. It is thus compact and suitable for fast evaluations. We performed an extensive architecture search to find the adopted hyperparameters.

4. RESULTS

In this section, we first assess the reliability of the Neural SI-SNR Estimator on synthetic data (using LibriMix and WHAMR!). Then, we will show the results achieved with the REAL-M dataset.

4.1. Results on LibriMix and WHAMR!

Figure 2 shows the scatter plots of the ground truth SI-SNRs and the estimated ones for the *single* training strategy. In this figure, we evaluate the SI-SNR estimator on the test sets of LibriMix and WHAMR! datasets when processing the mixtures with the SepFormer [1].

Interestingly, we notice a strong correlation. The Pearson coefficient is around 0.8 for both datasets. To ascertain whether the SI-SNR estimator works with different separation models, we tested it with mixtures processed by DPRNN and ConvTasNet (CTN). Note that the estimator is trained with the SepFormer only when adopting the *single* training strategy. Figure 3 shows the scatter plots under this mismatched condition. We can see that, even when tested with different separator models, the SI-SNR estimation correlates well with the oracle values. The Person correlation is still around 0.8 in all mismatched conditions.

In Table 1 we compare the *single* and *pool* training strategies described before. As expected, we observe a larger Pearson correlation with the *pool* strategy, which confirms the intuition that the SI-SNR estimator robustness is improved when training it with an ensemble of separators. For the rest of this work concerning real-world data, we use this SI-SNR estimator. We also note that the average absolute error on the test-set of WHAMR! is 1.54 dB, and 1.84 dB on LibriMix.

The results discussed in this section provide a preliminary indication of the effectiveness of the proposed approach. In the next section, we will extend the analysis to real-life mixtures.

4.2. Results on REAL-M Dataset

4.2.1. Subjective Opinion Scores

REAL-M is a dataset of real-life mixtures, and therefore no oracle SI-SNR is available. To validate the predictions of the neural estimator on real-world mixtures, we conducted a user study. We gathered the opinions of participants by asking them to assign a score between 1 (*'bad'* separation) and 5 (*'excellent'* separation) to the estimated sources.

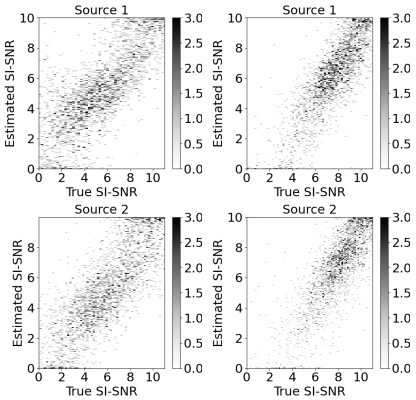

Fig. 2. Correlation plots for the SI-SNR estimates vs the ground truth SI-SNR values on the LibriMix dataset (left) and the WHAMR! dataset (right).

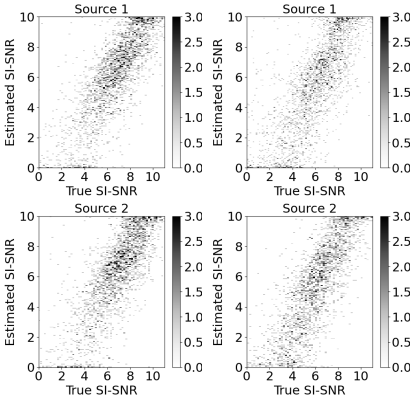

Fig. 3. Correlation plots for the SI-SNR estimates vs the ground truth obtained with Dual-Path RNN (left) and Convtasnet (right) on the WHAMR! dataset (mismatched conditions)

Table 1. Pearson Correlation Values (averaged over sources) for Different Separators on LibriMix and WHAMR! datasets, with two different SI-SNR estimators.

	SI-SNR-Estimator 1 *(single)*		SI-SNR-Estimator 2 *(pool)*	
Model	**LibriMix**	**WHAMR!**	**LibriMix**	**WHAMR!**
SF	0.80	0.81	0.82	0.87
DPRNN	0.80	0.80	0.83	0.84
CTN	0.81	0.79	0.85	0.86

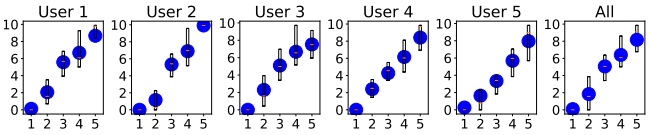

Fig. 4. A user study on estimated SI-SNR vs human opinion. The x-axes represent the participant preference (1 is 'bad' quality, while 5 is an 'excellent' quality). Y-axes represent the estimated SI-SNR. Blue Circles show the mean values, and the bars represent the 25th-75th percentile of each distribution.

For this study, we selected a subset of 50 mixtures and the corresponding estimated sources obtained with SepFormer trained on WHAMR!. We chose the mixtures to have a uniform distribution of estimated SI-SNR values between 0-10 dB. Figure 4 shows the scores obtained from five participants along with the aggregated one (rightmost plot). Each subplot shows the distribution of the estimated SI-SNR values for each of the opinion scores.

We observe that, on average, the estimated SI-SNR values are highly correlated with user opinions. Despite some variability, the average SI-SNR estimations linearly match with the human scores. In particular, the 25th-75th percentile of the distributions (shown with error bars around the mean of each distribution) correlates nicely with the user opinion scores. This result suggests that, on average, the neural estimator provides reliable predictions on the REAL-M dataset as well.

4.2.2. Estimated SI-SNR over training epochs

We now investigate how the estimated SI-SNR changes over the training epochs. Figure 5 shows the curves obtained when training the SepFormer [1], Dual-Path RNN [3], and Convtasnet [5] from scratch on the WHAMR! dataset. Figure 5 (left) shows the SI-SNR estimated by the proposed network when using REAL-M datasets as a validation set. Figure 5 (right), instead, shows the SI-SNR computed using the simulated WHAMR! validation dataset.

The training curve observed with the REAL-M dataset coupled with the proposed neural estimator, as expected follows a standard logarithmic trend for all the models. The performance improvement is larger in the first epochs, while a diminishing return is observed as long as training proceeds. Moreover, the models evaluated on REAL-M with the neural estimator achieve the same performance ranking obtained on the validation set of WHAMR!. In both cases, the best model is the SepFormer, followed by DPRNN and Convtasnet. This agreement is another indication of the reliability of the proposed estimator.

4.3. ASR based evaluation

The REAL-M dataset also provides the text transcription of each signal in the mixtures. Beyond using our blind SI-SNR estimator, it is thus possible to compute the Word-Error-Rate (WER) on the estimated sources. This metric can give another hint on the quality of the separator.

Table 2 reports the average WER using different speech recognizers and separators. In detail, we adopted a CRDNN model [24] trained on the LibriSpeech (LS) dataset [25] and a Wav2Vec 2.0 (W2V2) [27] model finetuned on the Common Voice (CV) English dataset [28]. Both models are implemented with SpeechBrain [24].

We observe that the Wav2Vec 2.0 model outperforms the CRDNN model. This result further confirms the relative effectiveness of Wav2Vec 2.0 even in challenging acoustic conditions [29].

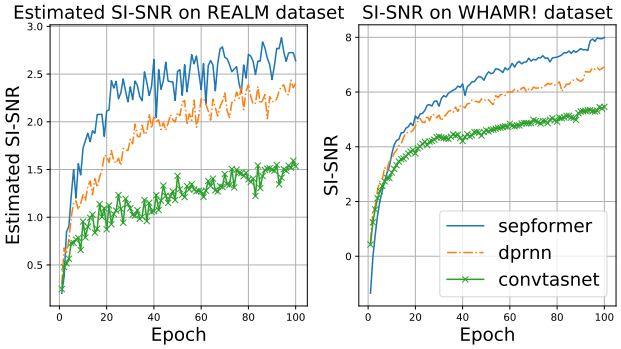

Fig. 5. Estimated SI-SNR vs training epochs for three popular source-separation models on the REALM dataset. (left) Achieved SI-SNR on the validation set of WHAMR! at the same epoch as left figure (right).

Table 2. Word-Error-Rates achieved on REAL-M with different speech recognizers and separators. The last column reports the estimated SI-SNR values.

Separator	**W2V2-CV** ↓	**CRDNN-LS** ↓	$\widehat{\textbf{SNR}}$ ↑
SF-WHAMR!	60.7	77.3	2.88
DPRNN-WHAMR!	64.9	78.4	2.43
CTN-WHAMR!	72.8	85.8	1.59

The absolute speech recognition performance, however, is still poor. The high WER highlights one more time how challenging speech separation is under real-life conditions. This is reasonable since the REAL-M mixtures are recorded in real environments in challenging and diverse acoustical conditions. Also in this context, we observe the same performance ranking between the achieved WERs and the SI-SNR estimations of the proposed SI-SNR estimator.

5. CONCLUSION

In this work, we release REAL-M, a dataset for speech separation in real-life settings, obtained through crowd-sourcing. We showed that a neural SI-SNR blind estimator can enable reliable evaluation of in-the-wild speech mixtures, for which the oracle target clean sources are unavailable. We extensively tested this approach, and we observed that the estimated SI-SNR values generally correlate well with the oracle SI-SNR values (on synthetic data) and with human assessments (on the real-life mixtures of REAL-M).

This study also highlights how challenging speech separation is on in-the-wild data. It is worth mentioning that speech separation in clean conditions (e.g., on WSJ0-2Mix) reaches 20 dB of SI-SNR with the best models for separation. The performance goes down to 8 dB when using simulated data with noise and reverberation (e.g., WHAMR!). It dramatically falls to 2.9 dB when using the real-life mixtures of REAL-M. This showcases that speech separation under real-life conditions is still an open problem, and we hope that our contribution will foster further research in this area.

We envision that, as future work, it is reasonable to explore the use of performance metrics other than SI-SNR, from speech enhancement literature [30, 31]. It is also a natural next step to test unsupervised approaches such as [32] on the REAL-M dataset.

6. REFERENCES

[1] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. of ICASSP*, 2021.

[2] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[3] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. of ICASSP*, 2020, pp. 46–50.

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

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[6] J. Chen, Q. Mao, and D. Liu, "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," in *Proc. of Interspeech*, 2020.

[7] Zhong-Qiu Wang, Ke Tan, and DeLiang Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *Proc. of ICASSP*, 2019.

[8] E. Nachmani, Y. Adi, and L. Wolf, "Voice Separation with an Unknown Number of Multiple Speakers," *Proc. of ICML*, 2020.

[9] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. of ICASSP*, 2016.

[10] P. Douglas and J. Baker, "The design for the wall street journal-based csr corpus," in *Proc. of HLT*, 1992.

[11] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step Sound Source Separation: Training on Learned Latent Targets," in *Proc. of ICASSP*, 2020.

[12] G. Wichern, Jo Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. of Interspeech*, 2019.

[13] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and Reverberant Single-Channel Speech Separation," in *Proc. of ICASSP*, 2020.

[14] Shinji Watanabe, Michael I. Mandel, Jon Barker, and Emmanuel Vincent, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *CoRR*, vol. abs/2004.09249, 2020.

[15] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus," in *Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen: Noldus Information Technology*, 2005.

[16] Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, Jian Wu, Xiong Xiao, and Jinyu Li, "Continuous speech separation: dataset and analysis," 2020.

[17] Maarten Van Segbroeck, Ahmed Zaid, Ksenia Kutsenko, Cirenia Huerta, Tinh Nguyen, Xuewen Luo, Björn Hoffmeister, Jan Trmal, Maurizio Omologo, and Roland Maas, "DiPCo — Dinner Party Corpus," in *Proc. Interspeech 2020*, 2020, pp. 434–436.

[18] Matthew Maciejewski, Shinji Watanabe, and Sanjeev Khudanpur, "Speaker Verification-Based Evaluation of Single-Channel Speech Separation," in *Proc. Interspeech 2021*, 2021, pp. 3520–3524.

[19] M. Yu, C. Zhang, Y. Xu, S.-X. Zhang, and D. Yu, "MetricNet: Towards Improved Modeling For Non-Intrusive Speech Quality Assessment," *CoRR*, vol. abs/2104.01227, 2021.

[20] C. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *CoRR*, vol. abs/2010.15258, 2020.

[21] P. Manocha, B. Xu, and A. Kumar, "NORESQA - A framework for speech quality assessment using non-matching references," *CoRR*, vol. abs/2109.08125, 2021.

[22] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement," in *Proc. of ICML*, 2019.

[23] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," in *Proc. of Interspeech*, 2021.

[24] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J. Chou, S. Yeh, S.-W. Fu, C.-F. Liao, E Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[25] 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.

[26] 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.

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

[28] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[29] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proc. of NeurIPS*, 2020.

[30] 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 *Proc. of ICASSP*, 2001, pp. 749–752 vol.2.

[31] L. Lightburn and M. Brookes, "A weighted stoi intelligibility metric based on mutual information," in *Proc. of ICASSP*, 2016, pp. 5365–5369.

[32] S. Wisdom, E. Tzinis, H. Erdogan, R. J Weiss, K. Wilson, and J. R Hershey, "Unsupervised sound separation using mixture invariant training," *Proc. of NeurIPS*, 2020.