INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Speaker Verification-Based Evaluation of Single-Channel Speech Separation

Matthew Maciejewski[1,2], *Shinji Watanabe*[1,2,3], *Sanjeev Khudanpur*[1,2]

[1]Center for Language and Speech Processing, The Johns Hopkins University, USA
[2]Human Language Technology Center of Excellence, The Johns Hopkins University, USA
[3]Language Technologies Institute, Carnegie Mellon University, USA

matt@mmaciejewski.com, shinjiw@ieee.org, khudanpur@jhu.edu

Abstract

Speech enhancement techniques typically focus on intrinsic metrics of signal quality. The overwhelming majority of deep learning-based single-channel speech separation studies, for instance, have relied on a single class of metrics to evaluate the systems by. These metrics, usually variants of Signal-to-Distortion Ratio (SDR), measure fidelity to the "ground truth" waveform. This can be problematic, not only for lack of diversity in evaluation metrics, but also in cases where a perfect ground truth waveform may be unavailable. In this work, we explore the value of speaker verification as an extrinsic metric of separation quality, with additional utility as evidence of the benefits of separation as pre-processing for downstream tasks.
Index Terms: speech separation, speaker verification

1. Introduction

In recordings of speech with multiple talkers present, such as conversational speech, it is common that people will speak at the same time, leading to overlapping speech [1, 2]. Most speech technologies suffer degradation of performance on overlapping speech [3], as they are designed to work on only a single speaker's speech. Speech separation aims to address this problem, producing a separate waveform for each speaker within a mixture, each containing only a single speaker's speech.

These conversational settings are often captured with room microphones, such as a table microphone in a meeting or a digital voice assistant in a home. With these generally far-field recordings, the increased speaker-to-microphone distance leads to decreased signal-to-noise ratio (SNR) for the speech in terms of both noise and reverberation [4]. The research community has accordingly begun to focus on these challenges, as performance in clean conditions has reached new heights [5, 6].

Training and evaluation of speech separation systems in noise and reverberation pose a number of challenges. One of the major challenges is that speech separation evaluation is generally done by computing the divergence of the estimated waveform from the ground truth waveform, as discussed in Section 2.1. The evaluation metrics are accordingly sensitive to all components of the waveform, penalizing performance for reasons other than *separation* errors, i.e. failing to produce the desired speech or failing to remove the other speakers. And, beyond this, these metrics cannot be used to evaluate *naturally* overlapped speech at all, as the ground truth separation waveforms are not available in any capacity.

As a result, we have conducted an investigation demonstrating the viability downstream speaker recognition as a method for evaluation of the quality of speech separation. Speaker recognition is the class of tasks aiming to identify the person who was speaking in a recording. If these systems are designed to only work on non-overlapping speech, we could expect that their performance would degrade as more and more of another speaker is present in a waveform, or—reframed in the context of speech separation—as the quality of separation is reduced. Speaker recognition has the benefit of being very lightweight with respect to annotation: rather than requiring an entire ground truth waveform, it requires only knowledge of the speakers identities. And, state-of-the-art speaker recognition systems are relatively invariant to noise and reverberation [7].

An additional benefit of this work is that it provides some evidence of the value in speech separation as pre-processing for speaker verification tasks. Many of the applications desired in conversational settings are either speaker recognition tasks or can be related to them, e.g. speaker diarization.

2. Discussion of Separation Evaluation

The methods of evaluation for system performance on a given task fall into two categories: direct evaluation and evaluation through down-stream task performance. In direct evaluation, some kind of performance metric is used to evaluate how closely system output matches the desired output. In evaluation through down-stream tasks, the system output is used as pre-processing for another system used on a different task, which is then evaluated.

2.1. Direct Separation Evaluation

2.1.1. Commonly-Used Metrics

The overwhelming majority of deep learning-based speech separation systems are evaluated using Signal-to-Distortion Ratio (SDR) metrics [8]. More recently, this means Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [9], but previously used the original SDR along with its less-reported companion metrics, Signal-to-Interferences Ratio (SIR) and Signal-to-Artifacts Ratio (SAR) [8] that use filtering and decomposition to classify errors. These metrics compute the ratio of signal to error in decibels, with error computed at the waveform level.

While rarely reported and not directly designed for speech separation, two additional metrics used are Short-Time Objective Inteligibility (STOI) [10] and Perceptual Evaluation of Speech Quality (PESQ) [11]. Like the SDR family of metrics, STOI and PESQ require access to the ground truth waveform.

2.1.2. Challenges of Direct Evaluation

One of the biggest challenges in direct evaluation of speech separation quality is that the level of ground truth annotation required is beyond the scope of something recoverable from any sort of natural mixture. For example, in a task such as automatic speech recognition (ASR), the annotation required for evaluation is simply the corresponding text. Systems can easily be trained and evaluated in noisy and reveberant conditions, as

http://dx.doi.org/10.21437/Interspeech.2021-1924

the labels themselves are unaffected by the audio condition. In contrast, speech separation evaluation struggles in these sorts of conditions, as it is not possible to recover the "clean" speech signal from the interfering signals. And, regardless of if the ground truth signal is clean or not, the performance metric will include the non-separation errors of failing to remove or produce noise in the estimates accordingly. There is also no way to evaluate separation performance on real, naturally-occurring speech mixtures—recovering the ground truth waveforms is itself the problem speech separation aims to solve.

Finally, as speech separation is a task largely desired as preprocessing for downstream speech technologies that are not designed for overlapping speech, a downside to direct evaluation is that the evaluation metrics are not guaranteed to correlate with the impact of the separation system on the performance of the downstream task.

2.2. Separation Evaluation Through Downstream Tasks

By considering evaluation through downstream tasks, there are considerably more options for performance evaluation, which additionally can be valuable for cases where the downstream task is the ultimate goal. However, a significant downside is that this approach is heavily dependent on the task, technique, and model used. And, in some cases, deep neural network (DNN)-based speech systems have been shown to have degraded performance on audio that has been produced or enhanced by a DNN [12, 13]. The most commonly desired downstream applications are speech recognition, speaker identification, and embedding-clustering–based diarization. Since the embeddings used in these systems typically come from speaker identification systems, we only consider the first two applications.

2.2.1. Separation Evaluation Through Speech Recognition

Though speech recognition of overlapping speech has largely been approached through end-to-end systems [14, 15], there has been some precedent of evaluating separated speech with speech recognition [16–18]. However, using ASR for downstream evaluation of speech separation does have downsides.

The biggest downside is that the data must contain complete utterances, which disallows the *min* condition defined in the wsj0-2mix dataset [19] and the use of any corpora that do not contain the constraint that the single-speaker waveforms contain full utterances [20]. In cases where both speech signals are full utterances, this almost assuredly leads to a condition that does not consist of 100% overlap, which is its own challenge in speech separation. And in cases of natural overlap, only a small portion of the utterances contain overlap. And finally, transcription is one of the more intensive and costly annotation procedures, particularly in comparison to speaker identity.

2.2.2. Separation Evaluation Through Speaker Verification

The biggest appeal of speaker identification-based evaluation is that speaker identity is very easy annotation to attain compared to the transcript required for ASR and ground truth waveform for direct evaluation. In addition, the annotation labels are time-invariant, applying to any duration of speech, avoiding many of the issues of ASR regarding segmentation and level of overlap and potentially serving as a valuable complementary metric for downstream evaluation.

There are, however, questions on how to evaluate the system. The primary decision is whether to perform speaker identification (speaker identity classification) or speaker verification (acceptance/rejection of the presence of a speaker from an enrollment recording), and how to extend those tasks to multi-speaker mixtures. For sake of simplification, in this work we focus on a verification task, where a trial is 'target' if the enrollment speaker is present in the mix and 'non-target' if none of the speakers in the mix are the enrollment speaker.

3. Experimental Configuration

3.1. Data

For our experiments, we used the wsj0-2mix dataset [19] based on WSJ0 [21], the WHAMR! dataset [6] that extends wsj0-2mix to noisy and reverberant conditions, the "noisy oracle" (no-2mix) dataset described in [22], as well as the mx6-2mix and ch5-2mix datasets described in [20] based on the Mixer 6 [23] and CHiME-5 [24] corpora respectively. In all cases, the 16 kHz versions and 'min' conditions are used. The wsj0-2mix corpus was chosen due to its ubiquitous use in nearly every deep learning-based speech separation study. WHAMR! was chosen due to the value of its conditions in many real-world applications. The no-2mix dataset was chosen to evaluate speaker verification as a solution to the evaluation challenges raised in its work. And finally, the mx6-2mix and ch5-2mix datasets were chosen for the purpose of evaluation in realistic environments, in contrast to the other datasets which contain fully-synthetic mixtures, with noise and reverberation being added to clean speech recordings after the fact.

For creating speaker verification test conditions, we chose to devise an algorithm for generating self-contained trials from the separation dataset evaluation conditions. In other words, trials are generated such that mixtures are paired with a single-speaker utterance that comes from the ground truth single-speaker recordings from other mixtures within that test set. This has a benefit of requiring no additional data from the source corpora and being easily applicable to all datasets, as they all share comparable structure and size. The algorithm selects utterances for a mixture based on constraints meant to maximize the diversity of speaker comparisons and variety of utterance usage.

To generate each speaker verification evaluation condition, we generated 2 target trials and 2 non-target trials for each mixture—the target trials consist of one speaker match for each of the two speakers in the mixture, and the non-target trials simply use a speaker that is different from both present in the mixture. It is worth noting that the trials we generated are not necessarily gender-balanced, instead approximately matching the gender balance of the source corpora. This is not ideal, but we felt the best option was to compromise on gender balance and focus on matching the conditions between the separation and speaker verification evaluation setups for a given dataset. In cases where separation trials correspond across datasets, the same trials were reused (wsj0-2mix, WHAMR!, and no-2mix use the same source recordings, and both mx6-2mix and ch5-2mix have the same mixtures across multiple microphones).

3.2. Networks

All of our separation networks are TasNet-BLSTM [25] networks with 600 units in each direction, trained with negative SI-SDR [9] loss, which we feel is reasonably representative of standard speech separation techniques. For the analysis and synthesis bases, we used 500 filters with length 5 ms and shift 2.5 ms. To train weaker models for our experiments with varying system performance, we increased the filter size and stride, as we feel there is evidence in the research community that demon-

strates that these parameters correlate strongly with separation performance [26].

Models were trained for 100 epochs using 4 second segments using the Adam [27] algorithm with an initial learning rate of 0.001. The learning rate is decreased by a factor of two if the validation loss does not improve for three consecutive epochs. In addition, gradient clipping is performed with a maximum ℓ_2 norm of 5. All networks were trained with either negative SI-SDR loss in an utterance-level permutation-invariant manner [28].

The system we used for speaker identification was xvector speaker embeddings [7] with a Probabilistic Linear Descriminant Analysis (PLDA) [29] backend for producing scores between utterances. We used models trained for the Speakers in the Wild evaluation [30] as described in [31]. The models are trained on VoxCeleb 1 [32] and VoxCeleb 2 [33], augmented with noise, music, babble, and reverberation. While this system was not trained on any overlapping or multi-speaker speech, we felt this was a reasonable system to use for our speaker verification backend due to its strong performance record, its design for noisy and reverberant environments comparable to desired speech separation application environments, and the fact that it was designed for an application in which multi-speaker environments were a key aspect of the evaluation. In addition, as the primary goal of this work is to evaluate speech separation quality, we do not necessarily want to maximize the invariance of the speaker identification technique to overlapping speech. And, as this system is reasonably current, it still lends claims to the value of speech separation as pre-processing for speaker identification.

3.3. Evaluation

The metric we use to represent the standard speech separation performance evaluation is Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [9]:

$$\text{SI-SDR}(\hat{s}) \coloneqq 10 \log_{10} \frac{\|s\|^2}{\|s - \beta\hat{s}\|^2} \text{ for } \beta \text{ s.t. } s \perp s - \beta\hat{s} \quad (1)$$

This measures the ratio of signal power to error power, using a scaled version of the estimated source such that the error is orthogonal to the signal.

For the speaker verification experiments, we use the most common metric, Equal Error Rate (EER). In typical speaker verification systems, a pairwise score is computed for each trial. The threshold used for acceptance/rejection of a trial can be tuned per application to favor either false acceptances and false rejections. EER reports the error rate at the operating point where the percentage of false acceptances and false rejections are equivalent.

In our mixture-based speech separation experiments, we used a straightforward approach to handling trials with multiple speakers. In cases where no separation was performed, we simply scored the enrollment utterance against the mixture itself. In cases where we were evaluating separated mixtures, we scored the enrollment utterance against both separated waveform, using the closest score as the ultimate score for the trial.

For our baseline experiments, we provide a sense of performance floor and ceiling for a given dataset and separation task by performing speaker verification evaluation on both the input mixture and the oracle ground truth separated signal, which exist due to the synthetic nature of mixtures required for standard speech separation training and scoring.

Table 1: *Documentation of performance across multiple conditions. EER [%] represents an error rate where smaller is better, while SI-SDRi [dB] measures signal improvement where larger is better. The Mix and Oracle columns provide an expected performance floor and ceiling, evaluating unprocessed mixtures and ground truth separation respectively, while Sys. columns report the performance of a TasNet separation system.*

Dataset	Mix EER	Oracle EER	Sys. EER	Sys. SDRi
wsj0-2mix [19]	13.7	2.4	4.7	14.5
no-2mix [22]	20.9	2.6	15.7	12.9
no-2mix w/ noisy target	22.4	7.4	13.1	2.8
WHAMR! [6] rev.	16.6	2.6	11.4	9.4
WHAMR! rev. w/ rev. target	17.8	4.6	9.3	9.9
WHAMR! noise & rev.	20.8	2.6	19.5	9.3
mx6-2mix [20] near	19.0	5.8	12.9	9.2
mx6-2mix far	23.7	11.2	21.6	2.3
ch5-2mix [20] near	33.8	35.4	36.4	6.9
ch5-2mix far	37.4	31.9	37.5	0.4

4. Results and Discussion

4.1. Survey of Conditions

Our first set of experiments was to simply evaluate and document performance across a wide variety of condition, the results of which are shown in Table 1. The purpose of these experiments were to document the potential gains for speaker verification through separation of overlapping speech, as well as show how close to the expected speaker verification performance ceiling a speech separation system is able to attain given its separation performance.

The results show that in general, there is a significant difference in performance of speaker verification between evaluating mixtures and evaluating the oracle separated speech, giving evidence that there is great potential for improvements to speaker recognition systems through separation pre-processing. One notable exception is the CHiME-5 conditions, where the lack of improvement is likely due to the conditions being exceptionally challenging. We also see from WHAMR! and no-2mix that although the speaker verification system was trained to be invariant to noise and reverberation, it tends to perform better on clean speech than noisy or reverberant speech.

In terms of the separation system performance, overall separation does generally improve performance over the baseline, with system EERs being lower than the EERs using the mixtures. One of the most interesting trends is in cases where the ground truth does or does not include the noise or reverberation (i.e., do we require the separation network to enhance the signal or not). In this case, separation performance is better when the network is asked to perform enhancement (drastically so in the no-2mix dataset), but speaker verification performance tends to do better when the network is not enhanced. Further discussion of this is in section 4.3.

4.2. System Comparison of SI-SDR to EER

Our next set of experiments were to investigate the relationship between SI-SDR and EER, with results shown in Figure 1. For these experiments, we varied performance of our separation system by changing the window size and shift for the TasNet bases.

The overall relationship among data points we collected is strictly monotonic, which is encouraging for the use of EER as a proxy metric for SI-SDR. We do, however, see the EER appear to level off at around double the oracle EER. This could sug-

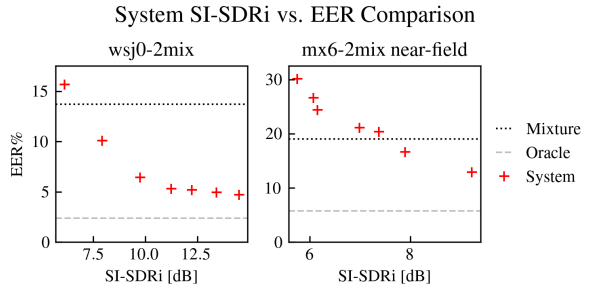

Figure 1: *Comparison between SI-SDRi and EER on wsj0-2mix and the near-field Mixer 6 condition over a variety of TasNet models with different performance attained with variable sliding window size and shift.*

gest a number of things: that better system verification performance would require very significant separation gains and that pushing separation performance higher has strongly diminishing returns; that there is some inherent limitation to this model and that the SI-SDR–EER curve is system-dependent; or some alternative explanation. Also it is worth noting that systems below a certain SI-SDRi around 7 dB performed worse on the speaker verification task than using the unprocessed mixture—even though those systems showed separation improvement over the mixture, the verification system performed worse using that separated output than using the mixture. This effect in wsj0-2mix, which contains no non-speech signals, supports the claim that the separation system is creating artifacts that the verification system is not robust to.

4.3. Noisy Ground Truth Results

Our final set of experiments were conducted to evaluate the sensitivity of the metrics to non-speech signals in the ground truth. The direct separation evaluation metrics are computed directly from the ground truth waveform and accordingly can penalize errors from parts of the signal that are not speech. Figure 2 shows the results of our experiments comparing the sensitivity of SI-SDR and EER to noise in the ground truth.

In all plots, both curves are the same two separation systems—the difference is that while they are trained using the same noisy mixtures, one is trained with clean sources and the other with noisy sources, as in [22]. Similarly, the left plots are evaluated according to clean source ground truth and the right noisy. The top plots report SI-SDR, the middle EER, and the bottom is EER using retrained, in-domain PLDA models.

A promising result is that the speaker verification results are consistent across both ground truth test conditions, and do show improvement in performance over the speaker verification baseline. In contrast, the direct separation evaluation differs greatly. And, not only does the performance differ, but the relative system ranking is inconsistent.

An unfortunate result is that the only result showing successful separation comes from the model trained with clean targets and evaluated with clean targets—which is not the better-performing system for speaker verification. A qualitative assessment of sample audio suggests that the clean-trained model produces high-quality separated and denoised audio, while the noisy-trained model produces separated-but-noisy speech, which suggests that perhaps the extra processing of the denoising of the clean-trained model may result in a greater amount of harmful DNN artifacts.

The third row represents experiments designed to address

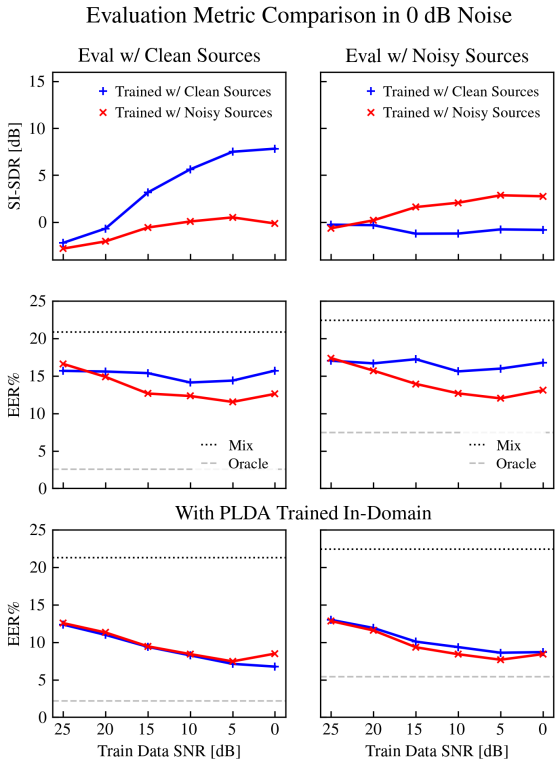

Figure 2: *Comparison between SI-SDRi and EER on a 0 dB no-2mix condition in both clean and noisy ground truth configurations. Note that larger numbers are better for SI-SDR while smaller are better for EER.*

this theory, using PLDA models trained using a combination of separated and enrollment utterances from the 'cv' set, with hopes that the PLDA could compensate for the effect of the artifacts. Interestingly, it not only improves performance but almost completely closes the gap between both systems. The speaker ID systems seem to be relatively invariant to the ground truth condition a separation network is trained on. This is encouraging for the use of separation as pre-processing, but adds further evidence for concerns about the use of SI-SDR as a metric in noisy conditions, as the first row suggests SI-SDR is very strongly related to the amount of noise in the signals.

5. Future Work

One continuation of this study we would like to explore is to extend the experiments to fully natural overlapping speech. This condition is the primary application for separation systems, but as of yet has not been evaluated outside of ASR. Another line of research we would like to pursue is a further investigation into the hypothesized impact of DNN artifacting on the verification system.

6. Conclusion

We have demonstrated the utility of speaker verification as a downstream evaluation of speech separation system performance, both through evidence of a monotonic relationship with SI-SDR and also through evidence of stronger invariance to non-speech signals present in the waveforms. Additionally, we have provided evidence that speech separation can improve performance of systems used in speaker recognition tasks.

7. References

[1] S. Bengio and H. Bourlard, *Machine learning for multimodal interaction*. Springer, 2005.

[2] Ö. Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition," in *Proc. ISCA Interspeech*, 2006.

[3] T. Yoshioka, I. Abramovski, C. Aksoylar, Z. Chen, M. David, D. Dimitriadis, Y. Gong, I. Gurvich, X. Huang, Y. Huang, A. Hurvitz, L. Jiang, S. Koubi, E. Krupka, I. Leichter, C. Liu, P. Parthasarathy, A. Vinnikov, L. Wu, X. Xiao, W. Xiong, H. Wang, Z. Wang, J. Zhang, Y. Zhao, and T. Zhou, "Advances in online audio-visual meeting transcription," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 276–283.

[4] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

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

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

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[8] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

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

[10] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[11] 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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752 vol.2.

[12] S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," in *Proc. ISCA Interspeech*, 2018, pp. 1571–1575.

[13] Z. Meng, J. Li, Y. Gong, and B.-H. F. Juang, "Adversarial feature-mapping for speech enhancement," in *Proc. ISCA Interspeech*, 2018, pp. 3259–3263.

[14] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6134–6138.

[15] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *arXiv*, 2021.

[16] T. Menne, I. Sklyar, R. Schlüter, and H. Ney, "Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech," in *Proc. ISCA Interspeech*, 2019, pp. 2638–2642.

[17] T. von Neumann, C. Boeddeker, L. Drude, K. Kinoshita, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR," in *Proc. ISCA Interspeech*, 2020, pp. 3097–3101.

[18] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, "Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party ASR," in *Proc. ISCA Interspeech*, 2019, pp. 1248–1252.

[19] J. R. Hershey, Z. Chen, and J. Le Roux, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 31–35.

[20] M. Maciejewski, G. Sell, Y. Fujita, L. P. Garcia-Perera, S. Watanabe, and S. Khudanpur, "Analysis of robustness of deep single-channel speech separation using corpora constructed from multiple domains," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2019, pp. 165–169.

[21] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," 1993, Web Download. Philadelphia: Linguistic Data Consortium.

[22] M. Maciejewski, J. Shi, S. Watanabe, and S. Khudanpur, "Training noisy single-channel speech separation with noisy oracle sources: A large gap and a small step," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[23] L. Brandschain, D. Graff, and K. Walker, *Mixer-6 Speech LDC2013S03*. Philadelphia: Linguistic Data Consortium, 2013.

[24] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. of Interspeech*, 2018.

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

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

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[28] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[29] S. Ioffe, "Probabilistic linear discriminant analysis," in *ECCV*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer-Verlag, 2006, pp. 531–542.

[30] M. McLaren, L. Ferrer, and D. Castán Lavilla, "The 2016 speakers in the wild speaker recognition evaluation," in *Proc. ISCA Interspeech*, 2016, pp. 823–827.

[31] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5796–5800.

[32] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proc. ISCA Interspeech*, 2017.

[33] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. ISCA Interspeech*, 2018.