MONAURAL SOURCE SEPARATION: FROM ANECHOIC TO REVERBERANT ENVIRONMENTS

Tobias Cord-Landwehr[⋆]*, Christoph Boeddeker*[⋆]*, Thilo von Neumann*[⋆]*,*
Cătălin Zorilă[†]*, Rama Doddipatla*[†]*, Reinhold Haeb-Umbach*[⋆]

[⋆] Paderborn University, Department of Communications Engineering, Paderborn, Germany
[†] Toshiba Cambridge Research Laboratory, Cambridge, United Kingdom

ABSTRACT

Impressive progress in neural network-based single-channel speech source separation has been made in recent years. But those improvements have been mostly reported on anechoic data, a situation that is hardly met in practice. Taking the SepFormer as a starting point, which achieves state-of-the-art performance on anechoic mixtures, we gradually modify it to optimize its performance on reverberant mixtures. Although this leads to a word error rate improvement by 7 percentage points compared to the standard SepFormer implementation, the system ends up with only marginally better performance than a PIT-BLSTM separation system, that is optimized with rather straightforward means. This is surprising and at the same time sobering, challenging the practical usefulness of many improvements reported in recent years for monaural source separation on nonreverberant data.

***Index Terms* —** speech separation, deep learning, SepFormer, automatic speech recognition, reverberation

1. INTRODUCTION

Neural network-based single-channel source separation has made significant advances in the last years. Starting with the seminal papers on deep clustering [1] and Permutation Invariant Training (PIT) [2], improvements have been achieved by combining the two in a multi-objective training criterion [3], or replacing the Short-Time Fourier Transform (STFT) with a learnable encoder and decoder [4]. Employing convolutional mask estimation network architectures [5] or accounting for short- and longer-term correlations in the signal with recurrent network layers [6] and combining them with a transformer architecture [7] further elevated the performance. Overall, this has led to an improvement in scale-invariant Signal-to-Distortion Ratio (SI-SDR) from roughly $10\,\mathrm{dB}$ to more than $20\,\mathrm{dB}$ on the standard WSJ0-2mix data set [1], which consists of artificial mixtures of nonreverberant speech.[1]

However, an anechoic environment is a rather unrealistic assumption for speech separation as in a real-world scenario, the superposition of the speech of two or more speakers typically occurs in a distant microphone setting. A distant microphone naturally captures a reverberated signal. A practically much more relevant setting is thus the separation of mixtures of reverberated speech.

Source separation of noisy and reverberant mixtures is much harder. In particular, reverberation has been considered more challenging than noise [8]. This comes to no surprise because the key assumptions underlying monaural mask-based source separation, namely the sparsity and orthogonality of speech representations in the STFT domain, tend to break down under reverberation.

WHAMR! [8] and SMS-WSJ [9] are two widely used data sets for research on source separation for reverberant mixtures. Both contain artificially reverberated utterances from the WSJ corpus. While WHAMR! additionally contains environmental noise, SMS-WSJ consists of 6-channel microphone array data and allows for performance comparison w.r.t. Word Error Rate (WER) as it is accompanied by a Kaldi recipe [10]. Source separation performance on WHAMR! is in the range of $2-8\,\mathrm{dB}$ output SI-SDR [2], while the performance on SMS-WSJ is in the range of $5-6\,\mathrm{dB}$ SI-SDR for single-channel input and single-stage processing [11, 8, 12], which is much worse than the performance on clean, anechoic mixtures. In this contribution, we employ SMS-WSJ for our experiments because we wish to assess the performance of the separation system not only by the signal-related evaluation metric Signal-to-Distortion Ratio (SDR) but also by WER, given that the SMS-WSJ Kaldi recipe allows us to compare the WER performance across different publications.

This paper is not about suggesting a new algorithm for reverberant source separation. We rather aim to explore, in a systematic way, which of the recent innovations that proved useful for the separation of anechoic mixtures are also beneficial in the reverberant case, in order to propose some guidelines on how to adjust a separation system to reverberated input.

As our outset, we take the SepFormer architecture, which achieves state-of-the-art performance both on WSJ0-2mix [7] and WHAMR! [13], and the traditional PIT-BLSTM source separation model from [2]. Here, we modify and optimize the PIT-BLSTM to detect which differences between both models aside from the separator lead to a better separation performance. Then, we modify the SepFormer w.r.t. loss function, encoder/decoder architecture and resolution to mitigate the performance degradation between the anechoic and reverberant scenario. Indeed, we are able to improve the performance w.r.t. WER by 7 percentage points compared to the vanilla SepFormer implementation. Nevertheless, the final result turns out to be hardly superior to that of the optimized PIT-BLSTM, calling into question the importance of some of the innovations of recent years for the realistic case of reverberant speech separation.

The remainder of the paper is structured as follows. In Section 2 the PIT-BLSTM and the SepFormer are briefly introduced as two realizations of an abstracted pipeline for mask-based source separation. Section 3 discusses design choices in light of the requirements of a reverberated input. In Section 4 the SepFormer is optimized for

[1] https://paperswithcode.com/sota/speech-separation-on-wsj0-2mix

[2] Obtained by comparing the reported improvement with the input SI-SDR of $-6\,\mathrm{dB}$

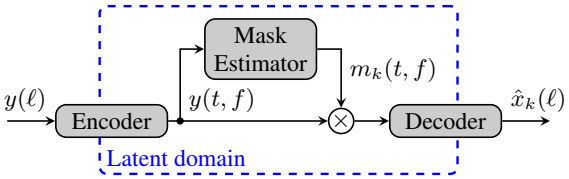

Fig. 1. Block diagram of mask-based source separation

performance on reverberant data and compared to the PIT model in Section 5. The paper concludes with a short discussion in Section 6.

2. MASK-BASED SOURCE SEPARATION

Mask-based systems for single-channel source separation can be abstracted to the same general processing pipeline that is depicted in Fig. 1. First, the observed time-domain signal $y(\ell)$ is transformed into a latent space (e.g., the STFT domain or a learned representation). In this latent space, the encoded mixture $y(t,f)$ with time index t and latent feature index f is used as the input of the neural separation module, which estimates a mask $m_k(t,f)$ for the reconstruction of each active speaker k in the observation. Then, the estimated signal $\hat{x}_k(t,f)$ of each speaker is obtained by masking the mixture with the estimated masks

$$\hat{x}_k(t,f) = y(t,f)m_k(t,f). \tag{1}$$

The reconstructed signals $\hat{x}_k(t,f)$ are then transformed back into the time-domain in the decoder.

Both the PIT-BLSTM approach to monaural source separation [2] and the SepFormer [7] (the latter providing state-of-the-art results on WSJ0-2mix) use a mask-based separation. By comparing these two models, which, in principle, share the same overall structure of Fig. 1, we investigate if modifications that were found to be useful in the anechoic scenario can be transferred to the reverberant case.

3. SOURCE SEPARATION UNDER REVERBERATION

Mask-based source separation relies on the sparsity and orthogonality of the sources in the the domain where the masks are computed. In case of the STFT domain, this means that a time-frequency bin (t,f) of a mixture $y(t,f)$ can be approximated by the contribution of the dominant source $i(t,f)$

$$\begin{aligned} y(t,f) &= \sum_{k=1}^{K} s_k(t,f)h_k(t,f) \\ &\approx s_{i(t,f)}(t,f)h_{i(t,f)}(t,f) \end{aligned} \tag{2}$$

where $s_k(t,f)$ and $h_k(t,f)$ are the STFT representations of the k-th source signal and the Room Impulse Response (RIR) from the k-th source to the microphone, respectively. Further, $i(t,f) \in \{1,\dots K\}$ indicates which of the K sources dominates in bin (t,f).

Note that Eq. (2) makes the additional assumption that the convolution of the source signal $s(\ell)$ with the RIR $h(\ell)$ corresponds to a multiplication of their respective STFT transforms. This so-called Multiplicative Transfer Function Approximation (MTFA), however, only holds true if the temporal extent of $h(\ell)$ is smaller than the STFT analysis window [14]. When the window length is decreased, this assumption becomes more and more questionable, and the Convolutive Transfer Function Approximation (CTFA) [15] would be more appropriate. Obviously, this challenges mask-based source reconstruction according to Eq. (1), and the complications are the more pronounced the smaller the STFT analysis window is.

When switching from a fixed STFT encoder to a learnable encoder, the overall structure of the system, see Figure 1, stays the same. Therefore, it can be assumed that similar issues arise with the learnable encoder. In the following we will thus study the influence of the encoder/decoder and their temporal resolution on the separation performance.

Table 1. Performance of the baseline models on the (anechoic) WSJ0-2mix database

Model	SDR	#Params
SepFormer [7]	20.4	25.7 M
SepFormer (small)	19.3	13.0 M
PIT-BLSTM [2]	9.8	23.5 M

Table 2. SDR of the baseline models on anechoic and reverberant SMS-WSJ data on the test dataset with matched training data

Model	anechoic		reverb	
	SDR	WER	SDR	WER
PIT-BLSTM	10.27	39.81	7.77	52.78
SepFormer (small)	**19.13**	13.14	8.98	41.43

4. EVALUATION

4.1. Database and Baseline Results

In order to assess which effect a specific component of a separation module has both on nonreverberant and reverberant data, it is important to run the experiments on a corpus that differs only in this respect. We employ the SMS-WSJ data set [9] for our analysis, which easily allows us to generate both anechoic and reverberant two-speaker mixtures that are identical aside from that.

For the anechoic scenario, the reverberation time T_{60} is reduced from 0.2 – $0.5\,\mathrm{s}$ to zero while keeping an otherwise identical data simulation. Dynamic mixing is employed in training: each example during training consists of randomly drawn utterances from WSJ database and only the RIRs are reused to provide a dramatically increased number of examples, which has been proven to improve the system's performance [12]. To show the competitiveness of the used models, we also provide baseline results on the WSJ0-2mix [1] database.

The PIT-BLSTM model consists of 3 BLSTM layers with 600 units each, followed by 2 fully connected layers. The encoder and decoder are set to the STFT and inverse STFT with a window size of 512, a frame advance of 128 and an embedding dimension (number of frequency bins) of 257 at $8\,\mathrm{kHz}$ sampling rate. The output of the STFT encoder is the concatenated real and imaginary part of the spectrum as in [16].

Aside from reducing the number of intra- and inter-Transformer layers to 4, we use the same SepFormer parameters as proposed in [7] with a window size of 16, a frame advance of 8 and a latent dimension of 256 samples.

This modification yields an about $1\,\mathrm{dB}$ lower SDR on WSJ0-2mix, but significantly reduces the number of parameters, see entry "SepFormer (small)" in Table 1. Thus, for all following experiments this "small" configuration is employed due to computational limita-

tions. Note that the memory footprint of the small SepFormer still is 16 times larger than the PIT-BLSTM, so that a complexity comparison purely based on the parameters is not fair. The learnable encoder is a single CNN layer with 256 channels, i.e. the latent size, followed by a ReLU, and the decoder has only one CNN layer as in [5].

Both architectures use the Adam optimizer [17] and the early reverberated signals as target as proposed in [9]. The SepFormer is trained with a soft-thresholded time-domain SDR loss [18]

$$\mathcal{L}^{\mathrm{th-SDR}} = 10\log_{10}\frac{1}{K}\sum_{k}\left(\frac{\sum_{\ell}|\hat{x}_k(\ell)-x_k(\ell)|^2}{\sum_{\ell}|x_k(\ell)|^2}+\tau\right), \quad (3)$$

where $\tau = 10^{-\mathrm{SDRmax}/10}$ and $\mathrm{SDRmax} = 20\,\mathrm{dB}$. This loss decreases the contribution of well separated examples to the gradient, encouraging the model to focus more on enhancing examples with a low SDR than those that already show a good separation. The Baseline PIT-BLSTM [2] is trained with a frequency-domain SDR loss. The models are evaluated w.r.t. SDR, PESQ [19], and WER. We use the SDR metric proposed in [20], as it allows an evaluation against the anechoic speech source. The PESQ values also are given w.r.t. the speech source, and the WER results on SMS-WSJ are determined with the acoustic model from [9].

Table 1 and Table 2 display the results of the baseline systems [2, 7] on WSJ0-2mix and SMS-WSJ, respectively. It can be seen that both systems degrade under the presence of reverberation. However, the separation performance of the SepFormer degrades by more than $10\,\mathrm{dB}$ in terms of SDR and almost 30 percentage points regarding the WER. We wish to find out which components of the SepFormer make it become so sensitive to reverberation.

4.2. PIT-BLSTM optimization

First, we optimized the performance of the PIT-BLSTM on reverberant input data. To do so, we switched the training objective from the frequency-domain loss to the thesholded time-domain loss described in Eq. (3). In this way, even though the PIT-BLSTM uses the magnitude spectrum for the mask estimation, the phase has an influence on the computed loss. In addition, we added white Gaussian noise at an SNR of $25\,\mathrm{dB}$ to the separated audio files before they were input to the speech recognizer. This is to mask artifacts that were introduced during the source separation. Besides an improved WER we also observed a higher correlation between the signal-level metric SDR and the WER, rendering the SDR a better predictor of the ASR performance. As shown in Table 3, by introducing the latter modifications the performance of the PIT-BLSTM is significantly improved both in terms of SDR and WER. Even more so, these modifications work well both with and without reverberation and lead to a reduction in WER of more than 20 percentage points for both scenarios.

4.3. SepFormer optimization

The above changes to the PIT-BLSTM system also lead to improvements of the SepFormer, see Table 3. Therefore, the Gaussian noise is added in all following evaluations. However, it is striking that the SepFormer is no longer superior to the PIT-BLSTM system for reverberant data. Therefore, we gradually exchanged the SepFormer's components with those of the PIT-BLSTM system to investigate the cause of this performance loss and what is the best configuration for reverberant input.

Table 3. Comparison of the optimzed PIT-BLSTM and the baseline SepFormer model on SMS-WSJ

Model	anechoic		reverb	
	SDR	WER	SDR	WER
PIT-BLSTM	10.27	39.81	7.77	52.78
PIT-BLSTM (th-SDR)	14.13	19.65	**10.93**	35.70
+ Gaussian noise	-	13.19	-	**27.47**
SepFormer (small)	**19.13**	13.14	8.98	41.43
+ Gaussian noise	-	**9.57**	-	33.51

4.3.1. Encoder/decoder choice

There is a large mismatch between the window size and the frame advance of standard PIT-BLSTM and SepFormer systems. To verify whether the violation of the MTFA caused by the small window size of the SepFormer contributes to the system deterioration under reverberation, we evaluated the SepFormer for multiple encoder/decoder configurations. As opposed to other works [16], we only increase the window size while maintaining small shift sizes in order to retain a high temporal resolution. Table 4 shows the expected behavior for the SepFormer in anechoic conditions: reducing the frame shift leads to an improvement in SDR and WER. The recommended analysis window size and shift of 16 and 8 samples (i.e. $2\,\mathrm{ms}$ and $1\,\mathrm{ms}$) [7], respectively, provides the best results for anechoic data. Furthermore, the learnable encoder proves superior to the STFT encoder.

Conversely, for the reverberant scenario, while the STFT encoder in Table 4 is significantly worse than a learnable encoder for a small window size and shift, it begins to be on par or even outperforms the learnable encoder for an increased window size of $32\,\mathrm{ms}$. This validates our assumption that the violation of the MTFA contributes to the poor model performance under reverberation. Interestingly, the overall best results of the SepFormer are achieved with the STFT.

Our assumptions are further supported by the W-Disjoint Orthogonality (WDO) [21] score which measures the orthogonality of the single-speaker utterances in the latent space. Following on the results from Table 4 it becomes apparent that the baseline SepFormer learns a highly orthogonal space for anechoic data. However, by switching to reverberant data, the WDO decreases by 5 percentage points. This decrease is mitigated by a larger encoder window size. The same is true for the STFT encoder, where the regularizing effect of a larger window size is even more pronounced. This indicates that the learnable encoder is able to compensate the effects to some degree, but that choosing a large enough window size is mandatory to stabilize the performance under reverberation.

4.3.2. Data representation for the mask estimator

A significant difference between PIT-BLSTM and SepFormer is that the PIT model estimates the masks based on the magnitude spectrogram only, whereas the SepFormer mask estimator has access to the complete signal, i.e., both magnitude and phase in case of the STFT representation.

To compare both networks with the same input representation, the effect of only using the magnitude as input for the mask estimator in the SepFormer is evaluated. The SepFormer trained with concatenated real and imaginary parts estimates separate masks for the real and imaginary parts of the observation, respectively. When only using the magnitude for the mask estimation, the estimated masks are

Table 4. Separation performance of the SepFormer on SMS-WSJ with a learnable and STFT encoder/decoder and varying encoder shifts/sizes

win. size	latent size	shift	learnable encoder	anechoic data				reverberant data			
				SDR [dB]	WDO [%]	PESQ	WER [%]	SDR [dB]	WDO [%]	PESQ	WER [%]
16	256	8	✓	**19.13**	**85.48**	**3.43**	**9.57**	8.98	79.94	1.83	33.51
256	256	8	✓	16.68	82.76	3.11	12.28	10.56	82.89	**1.91**	31.75
256	256	16	✓	15.27	81.89	3.00	13.96	10.23	81.95	1.85	30.66
256	256	64	✓	11.86	84.26	2.47	20.54	9.54	**85.08**	1.83	34.83
16	256	8	✗	16.74	73.58	2.84	11.90	7.44	69.95	1.71	45.83
256	256	8	✗	15.70	79.24	2.91	11.61	9.97	77.57	1.84	31.38
256	256	16	✗	14.47	79.24	2.69	13.52	**10.79**	77.57	1.90	**29.10**
256	256	64	✗	13.52	79.24	2.66	15.02	10.01	77.57	1.83	31.80

Table 5. Separation performance of the SepFormer for different input representations of the STFT features on reverberant SMS-WSJ

win. size	latent size	shift	Input data	SDR	PESQ	WER
256	256	16	Real+Imag	10.79	1.90	29.10
256	256	64	Real+Imag	10.01	1.83	31.80
512	256	16	Magnitude	10.48	1.82	29.22
512	256	128	Magnitude	**11.00**	**1.91**	**26.50**

Table 6. Performance comparison of the best anechoic and reverberant system configurations

System	anechoic		reverb	
	SDR	WER	SDR	WER
opt. PIT-BLSTM	14.13	13.19	10.93	27.47
opt. SepFormer anechoic	19.13	9.57	8.98	33.51
opt. SepFormer reverb	14.03	14.09	11.00	26.50

applied both on the real and imaginary parts. Table 5 shows that the availability of the phase information is not helpful for the SepFormer in the reverberant scenario. Even more so, omitting the phase information leads to a better system performance.

There are two possible reasons. Firstly, only using the magnitude spectrogram results in a larger window of 512 samples to keep the size of the separator identical, increasing the temporal context of each frame even further. Secondly, [22] has shown that the phase becomes less informative while the magnitude becomes more informative for increasing frame sizes. The configurations trained with both the phase and magnitude information learn a trade-off between phase and magnitude reconstruction. However, the large window sizes that were shown to be necessary in Table 4 for the reverberant scenario result in an uninformative phase representation. Therefore, omitting this information only slightly deteriorates the system performance for a small frame shift. However, by further increasing the frame shift the magnitude spectrogram becomes more informative. Therefore, using the magnitude allows increasing the frame shift from 16 to 128 samples, reducing the computational effort by almost a factor of 8 compared to the best configuration in Table 4 while simultaneously improving both signal-level metrics and WER.

5. SUMMARY

Table 6 summarizes the performance of the SepFormer on anechoic and reverberant SMS-WSJ using the best configuration for anechoic data as reported in [7] and the best configuration for reverberated input as found here, and compares it with the performance of the optimized PIT-BLSTM system. Interestingly, the SepFormer configuration that was found optimal for reverberant input is quite similar to the PIT-BLSTM: it uses a fixed STFT encoder with the magnitude spectrogram at its input and the same window size and frame shift. Only the network architecture of the separator is different, i.e., intra- and inter-transformer layers vs BLSTM layers. However, this modified SepFormer only shows a marginally better SDR and an improvement of 1 percentage point in the WER.

6. CONCLUSIONS

In this paper, we investigated the impact of reverberation on the various design choices for the SepFormer source separation system that is considered state-of-the-art for anechoic mixtures. We showed that it is mandatory to choose a large enough encoder window size for reverberant data. Also, we demonstrated that the STFT likewise is a valid choice as encoder and decoder. Here, it becomes apparent that the phase information no longer is helpful for the separation and only using the magnitude information provides superior results while reducing the computational complexity significantly.

Despite several modifications which greatly improved the performance of the SepFormer on reverberated mixtures, it was in the end hardly superior to a PIT-BLSTM separation system, which was optimized with only rather straightforward modifications, such as loss computation in time domain. At least for a single-stage approach, our experiments indicate that jointly focusing on phase and magnitude reconstruction leads to subpar results compared to solely focusing on magnitude reconstruction under reverberation. This raises the issue of whether the improvements that have been appraised for the separation of anechoic mixtures, such as learnable encoder and phase reconstruction, are futile for the more realistic case of reverberant source separation.

We therefore argue that research on source separation should primarily focus on the practically more relevant case of reverberant input, rather than on the anechoic scenario. Since jointly tackling both reverberation and overlapped speech appears to be a challenging task, an alternative solution is to apply an explicit dereverberation component and/or employ multi-stage processing, as in [23].

7. ACKNOWLEDGEMENT

Computational resources were provided by the Paderborn Center for Parallel Computing. C. Boeddeker was supported by DFG under project no. 448568305.

8. REFERENCES

[1] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[2] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker 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.

[3] Yi Luo, Zhuo Chen, John R. Hershey, Jonathan Le Roux, and Nima Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 61–65.

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

[5] Yi Luo and Nima 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.

[6] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.

[7] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 21–25.

[8] Matthew Maciejewski, Gordon Wichern, Emmett McQuinn, and Jonathan Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 696–700.

[9] Lukas Drude, Jens Heitkaemper, Christoph Boeddeker, and Reinhold Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

[10] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[11] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2001–2014, 2021.

[12] Neil Zeghidour and David 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.

[13] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Francois Grondin, and Mirko Bronzi, "On using transformers for speech-separation," *arXiv preprint arXiv:2202.02884*, 2022.

[14] Yekutiel Avargel and Israel Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[15] Ronen Talmon, Israel Cohen, and Sharon Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[16] Jens Heitkaemper, Darius Jakobeit, Christoph Boeddeker, Lukas Drude, and Reinhold Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6359–6363.

[17] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[18] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, and John Hershey, "Unsupervised sound separation using mixture invariant training," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 3846–3857.

[19] 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)*, 2001, vol. 2, pp. 749–752 vol.2.

[20] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[21] Scott Rickard and Ozgiir Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 1, pp. I–529–I–532.

[22] Tal Peer and Timo Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," *arXiv preprint arXiv:2203.16222*, 2022.

[23] Zhong-Qiu Wang, Gordon Wichern, and Jonathan Le Roux, "Convolutive prediction for monaural speech dereverberation and noisy-reverberant speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3476–3490, 2021.