On Data Sampling Strategies for Training Neural Network Speech Separation Models

William Ravenscroft[iD], Stefan Goetze[iD], and Thomas Hain[iD]

Department of Computer Science, *The University of Sheffield*, Sheffield, United Kingdom

{jwravenscroft1, s.goetze, t.hain}@sheffield.ac.uk

Abstract—Speech separation remains an important area of multi-speaker signal processing. Deep neural network (DNN) models have attained the best performance on many speech separation benchmarks. Some of these models can take significant time to train and have high memory requirements. Previous work has proposed shortening training examples to address these issues but the impact of this on model performance is not yet well understood. In this work, the impact of applying these training signal length (TSL) limits is analysed for two speech separation models: SepFormer, a transformer model, and Conv-TasNet, a convolutional model. The WJS0-2Mix, WHAMR and Libri2Mix datasets are analysed in terms of signal length distribution and its impact on training efficiency. It is demonstrated that, for specific distributions, applying specific TSL limits results in better performance. This is shown to be mainly due to randomly sampling the start index of the waveforms resulting in more unique examples for training. A SepFormer model trained using a TSL limit of 4.42s and dynamic mixing (DM) is shown to match the best-performing SepFormer model trained with DM and unlimited signal lengths. Furthermore, the 4.42s TSL limit results in a 44% reduction in training time with WHAMR.

Index Terms—speech separation, context modelling, data sampling, speech enhancement, transformer

I. INTRODUCTION

Speech separation models are used in a number of downstream speech processing tasks, from meeting transcription to assistive hearing [1]–[4]. Often, speakers are at a far-field distance from the microphone, which creates additional challenges for speech separation due to interference from noise and reverberation in the signal [5]–[7].

DNN separation models have led to significant improvements on anechoic data but there is a performance gap when these models are used for more distorted speech data [8]–[10]. Many of these models use Transformer or bidirectional long short term memory (BLSTM) layers [9]–[11] which can consume large amounts of memory and have quadratic time-complexity, i.e. for L input frames of data the model performs at least L^2 operations [12]. This is a particular concern in training when memory requirements are higher due to storing gradients for each operation required in the back-propagation stage [13]. This increased computational load also means longer training times. One way to compensate for the memory requirements is to use a batch size of 1 [9]

which leads to even longer training times as more parameter updates are performed. Another approach that reduces memory requirements and allows for larger batch sizes is to reduce the mixture signal length [14]. This reduces the training time but potentially at the expense of performance.

In this work, the first aim is to address if there are TSL limits at which no additional performance gain can be attained by DNN speech separation models. It is shown that, depending on model and dataset selection, there is a TSL limit at which not only no additional performance gain is attained but actually limiting the TSL to a specific value can lead to notably improved performance. This effect is demonstrated to be due to a random sampling of the start index when using TSL limits. Further evaluations show the benefit of having more unique training examples than using the full signal lengths. Finally, the application of TSL limits used with DM [15] is evaluated.

The remainder of this paper is structured as follows. In Section II, the signal model is introduced. The separation networks and datasets used are described in Section III, and Section IV, respectively. Section V presents evaluations for varying TSL limit for each separation network and dataset. Section VI explores splitting signals to generate more train examples and whether DM mitigates the effects gains found using TSL limits. Final conclusions are provided in Section VII.

II. SIGNAL MODEL

The noisy reverberant speech separation problem is defined as aiming to estimate C speech signals $\hat{s}_c[i]$ for sample index i and speaker number $c \in \{1, \ldots, C\}$ from the discrete time-domain mixture

$$x[i] = \sum_{c=1}^{C} s_c[i] * h_c[i] + \nu[i] \tag{1}$$

of length L_x. The $*$ operator denotes convolution, $h_c[i]$ is a room impulse response (RIR) corresponding to speaker c and $\nu[i]$ denotes additive background noise.

III. SEPARATION MODELS

The SepFormer [9] and Conv-TasNet [14] models are both widely researched time-domain audio separation networks (TasNets). The structure for these TasNets is to have a time-domain neural encoder which encodes a mixture signal block $\mathbf{x}_\ell$ of size L_{BL} to $\mathbf{w}_\ell$ followed by a mask estimation network which estimates a series of masks $\mathbf{m}_\ell$ for each of C speakers. These masks are used to separate the encoded features

which are then decoded back in to a time domain signal $\mathbf{s}_{\ell,c}$ using a neural decoder. An example of the architecture for $C = 2$ speakers can be seen in Figure 1. Both Conv-
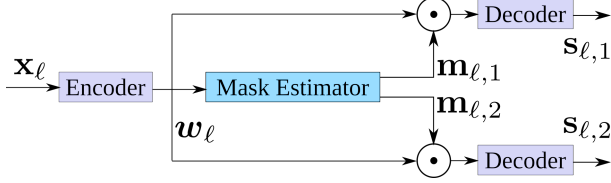

Figure 1. Architecture of the SepFormer and Conv-TasNet models, exemplary for $C = 2$ speakers. The $\odot$ symbol denotes the Hadamard product.

TasNet and SepFormer are trained using the utterance-level permutation-invariant scale-invariant signal-to-distortion ratio (SISDR) objective function [9], [14], [16], [17]. Models are trained according to the best performing models in each of their original papers [9], [14] unless stated otherwise. Batch sizes of 2 and 4 are used for SepFormer and Conv-TasNet respectively except where otherwise stated.

A. SepFormer Network

The SepFormer model is briefly introduced in this section. SepFormer is chosen because it is a large transformer model that is among the state-of-the-art models on several speech separation benchmarks [9], [18]. The SepFormer uses a 1D convolutional layer for encoding the signal proceeded by a rectified linear unit (ReLU) activation function. The decoder is a single transposed 1D convolutional layer. The mask estimation network uses a dual-path structure [11] whereby a series of Transformer layers are stacked such that each alternating layer computes a multihead attention layer on either the local or global context of the sequence. The local processing is achieved by first splitting the input signal into overlapping chunks of a predetermined size K turning a batched 3D tensor into a 4D tensor. The output of the local Transformer layer is a 4D tensor which is then reshaped by swapping the axes of encoded chunks and the number of chunks before being fed into the global Transformer layer. The final stage is to reconstruct a 3D tensor using the overlap-add method to produce C sequences of masks. The encoded features are then masked and decoded back into the time domain.

B. Conv-TasNet

The Conv-TasNet model contrasts the SepFormer in that it is a much smaller model (25.8M vs. 3.5M parameters in the implementations used here) and the only global information it processes is the overall signal energy, whereas the SepFormer model has global access to all information in the input signal due to the transformer layers used. Conv-TasNet uses a temporal convolutional network (TCN) sequence model for the mask estimator in Figure 1. The encoder and decoder of the network are the same as those used for the SepFormer model but with a different number of output channels in the configuration used in this paper. The mask estimation network is composed of a 1D pointwise convolution (P-Conv)

bottleneck layer, a TCN and a 1D P-Conv projection layer with a ReLU activation function to produce the sequence of masks. The TCN is composed of a series of convolutional blocks consisting of P-Conv and depthwise-separable convolution (DS-Conv) layers with kernel size P. The convolutional blocks are configured in stacks of X blocks with increasing dilation factors $f \in \{2^0, 2^1, \ldots, 2^{X-1}\}$. Stack are repeated R times with the dilation factor reset at the start of each stack [14].

IV. DATASETS

Three corpora of 2-speaker mixtures are analysed in this work. The trends demonstrated later in Sections V and VI for the 2-speaker scenario are assumed to generalize to higher C values. For all corpora, the 8kHz *min* configuration is used. The *min* configuration refers to mixtures being truncated to the shortest utterance in a mixture as opposed to padding shorter utterances to the longest utterances.

A. WSJ0-2Mix and WHAMR

WSJ0-2Mix and WHAMR are both simulated 2-speaker corpora derived from the WSJ0 corpus [19], [20]. WSJ0-2Mix takes speech samples from WSJ0 and overlaps them at speech-to-speech ratios (SSRs) between 0 and 5 dB. WHAMR is a noisy reverberant extension of WSJ0-2Mix with noise from the WHAM [21] dataset mixed with the loudest speaker at signal-to-noise ratios (SNRs) between -6 and 3 dB.

B. Libri2Mix

Libri2Mix is a simulated 2-speaker mixture corpus derived from the LibriSpeech and WHAM corpora [22]. Speech samples come from the LibriSpeech corpus [23] and noise samples come from the WHAM corpus. Instead of SSRs, LibriMix uses loudness units relative to full scale (LUFS) measured in dB to set the loudness of speakers and noise in the mixtures. Speakers have a loudness between -25 and -33 LUFS and noise has a loudness between -38 and -30 LUFS. For training, the *train-100* dataset was chosen as it has a very similar TSL distribution to the alternate *train-360* dataset but with less examples, meaning shorter training times.

C. Signal Length Distributions

The distribution, density estimation (DE) and the mean and standard deviation of the mixture signal length in the WHAMR train *tr* and test *tt* sets can be seen in Figure 2. WSJ0-2Mix and WHAMR have identical signal distributions as WHAMR is derived from the former. These distributions are shown in the left panel. The train and test sets in WHAMR have similar distributions of signal length with mean values within 0.3s of one another and standard deviations within 0.1s of one another. This contrasts the distributions of the Libri2Mix dataset [22] also shown in Figure 2 where the *train-100* and *test* sets have a difference in mean value of 6.2s and difference in standard deviation of 1.79s.
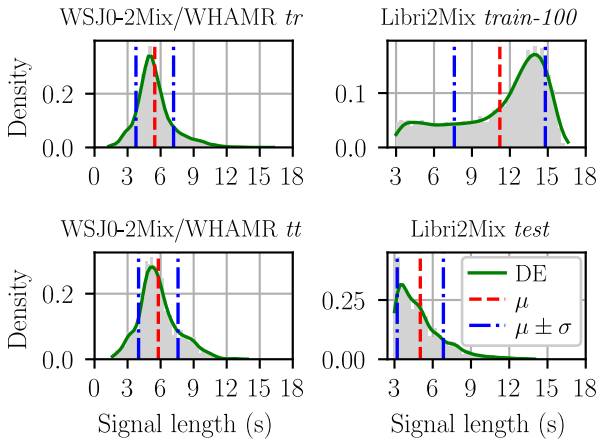
Figure 2. Distributions of mixture signal lengths in WSJ0-2Mix/WHAMR (left) and Libri2Mix (right) for both train (top) and test (bottom) sets. Density estimation (DE) is shown by solid green lines, mean values are indicated by dashed red lines and standard deviation values by dash-dotted blue lines.

V. TRAINING SIGNAL LENGTH ANALYSIS

Evaluations of varying the TSL limit are presented in this section. For all evaluations, the improvement in SISDR over the input mixture signal, denoted by Δ SISDR, is used as the evaluation metric. SISDR measures the energy of distortions in the estimated speech signals and is one of the most common metrics used in recent monaural speech separation literature [9]–[11].

A. Initial TSL Limit Evaluations

As a first experiment, twelve SepFormer models are trained and evaluated on WSJ0-2Mix, WHAMR and Libri2Mix, each with a different TSL limit. Twelve logarithmically spaced signal limits $T_{\lim}$ are selected between 0.5s and 10s:

$$T_{\lim} \in \{0.5, 0.66, 0.86, 1.13, 1.49, 1.95,$$
$$2.56, 3.36, 4.42, 5.8, 7.62, 10\}\text{s}. \quad (2)$$

The notation $L_{\lim}$ is used for the respective discrete sample index, i.e. $L_{\lim} = T_{\lim}f_s$ for sampling rate f_s. When cutting the training signal lengths such that $L_x \leq L_{\lim}$ the starting sample index of the signal is randomly selected from the uniform distribution $\mathcal{U}(0, 1 + max(0, L_x - L_{\lim}))$. Performance for SepFormer models trained and evaluated on all three datasets is compared in Figure 3. For the WHAMR corpus, an increase in overall Δ SISDR performance from the 0.5s to 1.95s limit can be observed. The optimal TSL is at 3.36s. Between 3.36s and 10s performance decreases again by -1.4dB. This may seem surprising as the general convention with training DNNs is that more data normally results in improved overall performance. A similar trend is observed for WSJ0-2Mix where there is a notable increase between 0.5s and 3.36s and then a drop in performance of 0.8dB between 4.4s and 10s. For Libri2Mix, the performance saturates before a TSL limit of 4.42s. There is no drop in performance as the TSL limit approaches 10s which is likely due to the Libri2Mix training set having a more uniform distribution below signal lengths of 10s than the WHAMR or WSJ0-2Mix datasets, cf. Figure 2.
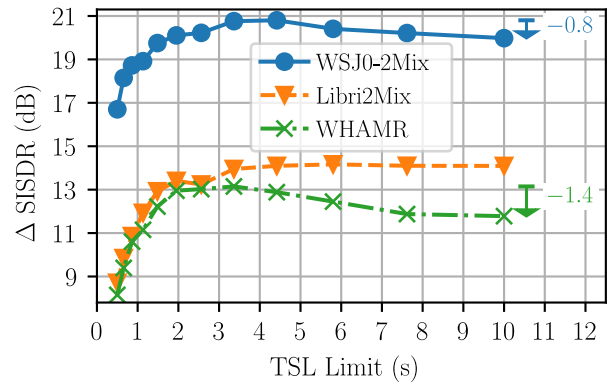

Figure 3. SepFormer results for varying the TSL limit for the anechoic WSJ0-2Mix (top), Libri2Mix (middle) and WHAMR (bottom) test sets.

The results for the WHAMR evaluation set are separated into quartiles of mixture signal length for the following experiment. Δ SISDR results for each quartile are shown in Figure 4. Comparing Q1 to Q4 shows that, with a sufficiently large
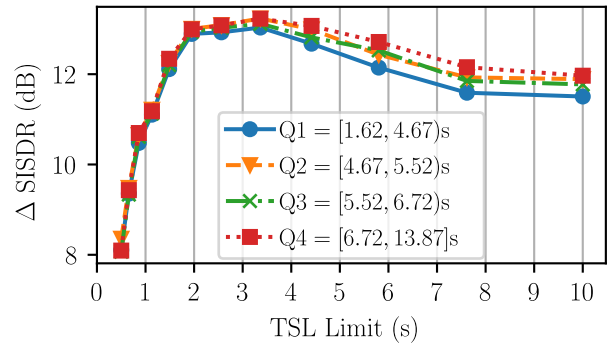

Figure 4. Training signal length (TSL) analysis of the 1st to 4th signal length quartiles in the WHAMR evaluation set.

TSL (≥ 1.95s), the best separation performance in SISDR is found on the longest signal lengths, regardless of TSL. A loss in SISDR performance is still observed from 3.36s to 10s regardless of which quartile is evaluated.

B. Training Time Evaluation

The average training epoch duration (ED) for the SepFormer model on WHAMR and Libri2Mix training sets are shown in Figure 5. Note the ED for WSJ0-2Mix is omitted for brevity but is similar to WHAMR due to having the same TSL distribution (cf. Figure 2). All models were trained on the same hardware to control any impact this has on speed. The average EDs for the WHAMR dataset have a sigmoidal shape due to the majority of the signal lengths being concentrated around the mean signal length of the training set (5.6s, cf. Figure 2). Libri2Mix has a more linear relationship between TSL limit and ED due to the more uniform shape of the signal length distribution below 10s in the *train-100* set, cf. Figure 2. Reducing the TSL limit has more benefit in terms of ED for the Libri2Mix dataset when iterating over all training examples.
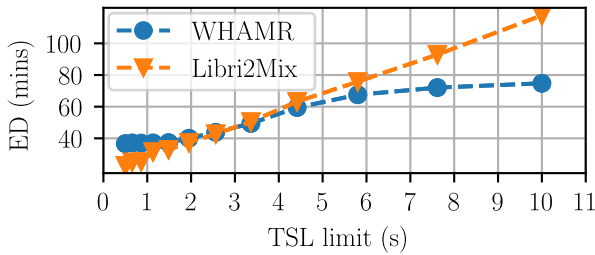
Figure 5. Comparison of average epoch duration (in mins) for the SepFormer model on the Libri2Mix and WHAMR training sets.

C. Fixed vs. Random Start Index

In Section V-A, the start index of each shortened signal was randomly sampled from a uniform distribution. In this section, this is compared to using a fixed start sample. A start sample of 1999 ($= 0.25$s at 8kHz) was used for signals where the original mixture signal length was larger than $L_{\lim}$, else the entire signal length was used. The motivation for this was that many training examples contain silence at the beginning of clips. It was considered desirable to omit as much silence to make for a fairer comparison with the randomly sampled clips which are assumed to have a lower likelihood of beginning with silence. Results in Figure 6 confirm that the loss in performance from a TSL limit of 3.36s to 10s with WHAMR is due to randomly sampling the start index. The
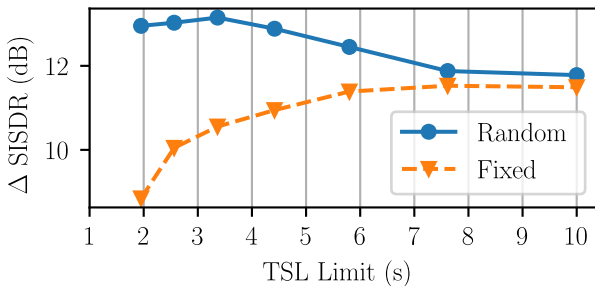

Figure 6. Comparison of TSL variation for the SepFormer model trained and evaluated on the WHAMR datasets on a subset of TSL limits in the range [1.95, 7.62]s.

performance saturates at a TSL limit of 5.8s when using a fixed start index. This is similar to the performance saturation point of Libri2Mix in Figure 3 demonstrating the performance drop in higher TSLs seen before on WHAMR (cf. Figure 3) is related to both a non-uniform TSL distribution and the random sampling used.

D. Transformer vs. Convolutional Model

Results comparing the Conv-TasNet model (cf. Section III-B) to the SepFormer model are shown in Figure 7. The loss in performance above 3.36s is not observed for the Conv-TasNet model. All SISDR results above $T_{\lim} = 1.95$s are within in 0.5dB of each other suggesting Conv-TasNet is more invariant to the TSL limit if the limit is sufficiently large. This is possibly due to the 1.53s receptive field of the Conv-TasNet models being smaller than these particular TSLs limits [24].
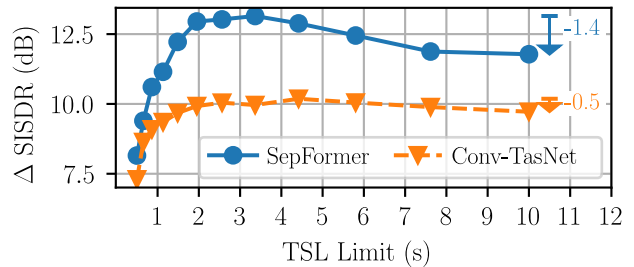

Figure 7. Comparison of SepFormer and Conv-TasNet across TSL limits $T_{\lim} \in [4.42, 7.62]$ using the WHAMR corpus.

VI. SIGNAL SPLITTING AND DYNAMIC MIXING

Next, two sampling strategies are evaluated to (i) investigate whether the performance gained by TSL with random sampling on shorter sequences still holds when the same quantity of audio data in terms of length in seconds is used and (ii) whether using TSL limits still result in performance gains with DM, i.e. simulating new speech mixtures for each epoch [15].

A. Signal Splitting

A signal splitting strategy was designed such that a batch of inputs $\mathbf{X} \in \mathbb{R}^{M \times L_x}$ was reshaped to $\mathbf{X}' \in \mathbb{R}^{MD \times \frac{L_x}{D}}$ for batch size M. Signal length L_x is still limited such that $L_x \leq L_{\lim}$. The motivation of this method is to evaluate the importance of training on the entire sequence length compared to the raw data quantity used in seconds. Computational complexity in training is also reduced. TSLs for $T_{\lim} \in [4.42, 10]$s are analysed for $D = 2$. Figure 8 shows that $D = 2$ improves performance for shorter TSL limits ($T_{\lim} \leq 5.8$s) compared to $D = 1$ (the original shape). However for $T_{\lim} \geq 7.62$s the performance is similar to $D = 1$. As in Figure 3 this is likely due to the TSL distribution of WHAMR.
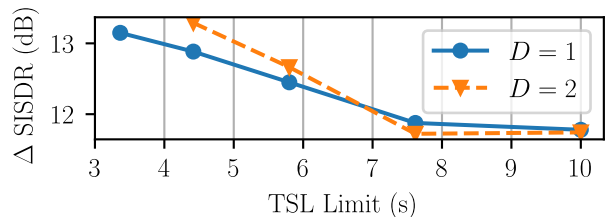

Figure 8. Comparison of split signal and batch reshape training $D = 2$ against full signal training $D = 1$ for the SepFormer model.

B. Dynamic Mixing

DM was proposed to improve performance of various speech separation models [15]. DM often results in a 1.0 to 1.5dB SISDR performance improvement dependant upon the model and dataset [15], [25]. The random start index sampling used in Section V is similar to DM in that it provides the model with unique training examples each epoch but without simulating new mixtures. In this section using DM and TSL limits is compared against just using TSL limits to see if further performance gains can be attained using both approaches. The DM results for the WHAMR corpus are

shown in Figure 9. It can be seen that with DM the drop in performance is less (at 7.62s) than without DM. The best
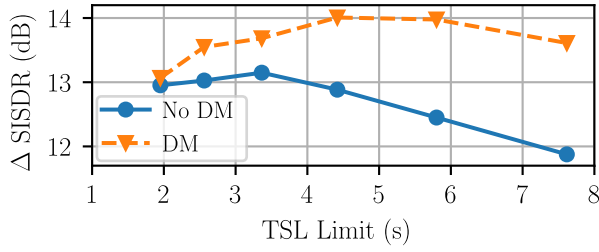

Figure 9. Comparison of SepFormer model training using TSL limits with and without dynamic mixing being used as well on WHAMR evaluation set.

performing model $T_{\text{lim}} = 4.4$s is compared to the Sepformer model with no TSL limit in Table I. The batch size reported for the model with no TSL is the largest it was found possible to train on a 32GB Nvidia V100 GPU. It can be seen that using the TSL limited model is able to match its performance with an average ED reduction of 44%, highlighting the benefit of this approach.

TABLE I
COMPARISON OF BEST PERFORMING SEPFORMER MODELS ON WHAMR WITH AND WITHOUT TSL LIMITS. M DENOTES BATCH SIZE AND THE AVERAGE EPOCH DURATION (ED) IS REPORTED IN MINUTES.

Model	M	T_{lim} (s)	$\triangle$ SISDR (dB)	ED (mins)
SepFormer + DM	1	—	**14.0**	85
SepFormer + DM	2	7.62	13.6	74
SepFormer + DM	2	4.42	**14.0**	**59**

VII. CONCLUSION

In this paper it was shown that TSL limits can affect the overall performance for speech separation models in a number of ways. For WSJ0 derived speech separation benchmarks, i.e. WSJ0-2Mix and WHAMR, it is optimal to use shortened training examples randomly sampled from the original examples due to the signal length distribution of these corpora. For the Libri2Mix dataset, the same method led to shorter training times with no notable loss in performance. The SepFormer model was compared to the Conv-TasNet model and it was shown that the Conv-TasNet performance has less variation. Using dynamic mixing and TSL limits with random sampling was shown to be able to match the performance of the SepFormer model trained DM using full sequence lengths on WHAMR with a 44% reduction in training time. With some previous literature opting to limit TSLs [11], [14] and others not [9], [10] on the same benchmarks, the results in this paper suggest that this is not a fair comparison and that TSL limiting is important to factor in when analysing results, particularly for the WSJ0-2Mix and WHAMR benchmarks.

REFERENCES

[1] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-Field Automatic Speech Recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2021.

[2] N. Moritz, K. Adiloğlu, J. Anemüller, S. Goetze, and B. Kollmeier, "Multi-channel speech enhancement and amplitude modulation analysis for noise robust automatic speech recognition," *Computer Speech & Language*, vol. 46, pp. 558–573, November 2017.

[3] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, N. Kanda, J. Li, S. Wisdom, and J. R. Hershey, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," *SLT 2021*, Jan 2021.

[4] B. J. Borgström, M. S. Brandstein, G. A. Ciccarelli, T. F. Quatieri, and C. J. Smalt, "Speaker separation in realistic noise environments with applications to a cognitively-controlled hearing aid," *Neural Networks*, vol. 140, pp. 136–147, 2021.

[5] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Interspeech 2018*, Sep. 2018.

[6] T. Cord-Landwehr, C. Boeddeker, T. von Neumann, C. Zorila, R. Doddipatla, and R. Haeb-Umbach, "Monaural source separation: From anechoic to reverberant environments," in *IWAENC 2022*, Sep. 2022.

[7] W. Ravenscroft, S. Goetze, and T. Hain, "Att-TasNet: Attending to Encodings in Time-Domain Audio Speech Separation of Noisy, Reverberant Speech Mixtures," *Frontiers in Signal Processing*, vol. 2, 2022.

[8] D. Ditter and T. Gerkmann, "Influence of Speaker-Specific Parameters on Speech Separation Systems," in *Interspeech 2019*, Sep. 2019.

[9] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Interspeech 2021*, July 2021.

[10] J. Rixen and M. Renz, "QDPN - Quasi-dual-path Network for single-channel Speech Separation," in *Interspeech 2022*, Sep. 2022.

[11] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020*, Oct 2019.

[12] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *ICML 2020*, Jul. 2020.

[13] S. Haykin, *Neural Networks and Learning Machines*, 10th ed. Prentice Hall, 2009.

[14] 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.

[15] 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.

[16] M. Kolbaek, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 10, p. 1901–1913, oct 2017.

[17] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *ICASSP 2019*, May 2019.

[18] C. Subakan, M. Ravanelli, S. Cornell, F. Grondin, and M. Bronzi, "On using transformers for speech-separation," 2022. [Online]. Available: https://arxiv.org/abs/2202.02884

[19] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-Channel Multi-Speaker Separation using Deep Clustering," in *ICASSP 2016*, Sep. 2016.

[20] M. Maciejewski, G. Wichern, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *ICASSP 2020*, May 2020.

[21] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending speech separation to noisy environments," in *Interspeech 2019*, Sep. 2019.

[22] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020. [Online]. Available: https://arxiv.org/abs/2005.11262

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP 2015*, April 2015.

[24] W. Ravenscroft, S. Goetze, and T. Hain, "Receptive Field Analysis of Temporal Convolutional Networks for Monaural Speech Dereverberation," in *EUSIPCO 2022*, Aug. 2022.

[25] ——, "Deformable temporal convolutional networks for monaural noisy reverberant speech separation," in *ICASSP 2023*, Jun. 2023.