COMBINING CONFORMER AND DUAL-PATH-TRANSFORMER NETWORKS FOR SINGLE CHANNEL NOISY REVERBERANT SPEECH SEPARATION

William Ravenscroft, *Stefan Goetze*, *and Thomas Hain*

Department of Computer Science, *The University of Sheffield*, Sheffield, United Kingdom
{jwravenscroft1, s.goetze, t.hain}@sheffield.ac.uk

ABSTRACT

Separation of overlapping speakers remains an active area of speech technology research. Many deep neural network (DNN) separation models propose modelling local and global temporal context separately using alternating DNN layers. Two such models are SepFormer and TD-Conformer. The largest configurations of each have comparable computational cost and similar performance; with SepFormer performing better on anechoic data and TD-Conformer yielding better results on noisy reverberant data. This work combines these two model types to gain insights into how their computational characteristics affect their performance. The generalization benefits of the larger model size of the conformer layers are demonstrated both on the WHAMR and the out-of-domain far-field evaluation set MC-WSJ-AV across a number of evaluation metrics. The proposed model is able to achieve 22.1 dB and 14.7 dB average scale-invariant signal-to-distortion ratio (SISDR) improvement when trained and evaluated on WSJ0-2Mix and WHAMR, respectively. The model trained using WHAMR is able to achieve 4.3 dB average SISDR improvement on the out-of-domain MC-WSJ-AV dataset.

Index Terms— speech separation, speech enhancement, neural networks, conformer, dual-path transformer

1. INTRODUCTION

Speech separation and related technologies such as speaker extraction are important for many real-world applications [1] such as digital assistants [2, 3], automatic meeting transcription [4, 5] and assistive hearing [6]. Recent speech separation research has heavily focused on time-domain audio separation network (TasNet), dual-path (DP) modelling and attention (or transformer) networks [7–10]. TasNet models, first proposed in [11], are typically composed of an encoder, a mask estimation or mapping network, and a decoder where the encoder encodes signals from the time domain using a neural network layer and the decoder decodes this neural representation back into the time-domain [12]. DP separation models, proposed in [13], use alternating neural network layers to process local and global contexts separately. One such DP transformer model, known as SepFormer [14], is one of the most performant models on separation benchmarks such as WSJ0-Mix and WHAMR [15, 16]. In this model, the input sequence is first split into fixed-size chunks which are input to a transformer layer for processing the local context. The *chunk size* and *number of chunks* axes are swapped and the Tensor is then processed by another transformer layer to model the global context. An analogue of the DP structure is the conformer model [9, 17, 18] where the local context is processed by a convolution module instead of a transformer. This approach generally has lower computational complexity for processing the local context for a fixed feature dimension but comes at the cost of increased model size. In the DP layers, the swapping of the axes, as opposed to processing the reconstructed feature sequence, reduces the computational complexity of the attention function in the global context layer. In the time domain conformer (TD-Conformer) model, proposed in [9], a subsampling layer is used to reduce the temporal resolution of the input sequence similarly and thus the computational complexity of the transformer layer used to process the global context.

In this work the convolutional separation transformer (ConSepT) model is proposed. ConSepT is a mixed Conformer and DP transformer model. The motivation for combining the two variant layers is that controlling for model complexity DP transformer models have been shown to be more performant for anechoic speech mixtures and conformer models have been shown to be more performant on noisy reverberant speech mixtures [9, 14]. This contrast is explored by mixing the two layer types to analyse if a higher overall performance can be obtained by combining the two. The model is structured so that conformer layers process the earlier features in the network under the assumption that they contain more noise and thus the DP transformer layers are used to process the cleaner features. There are two key contrasts between the two model types; firstly, the conformer layers result in a larger model size primarily due to the convolutional local context layer in the conformer block, and secondly, the DP transformer layers typically have significantly more computational complexity given to processing local context whereas conformer layers give most of their computational complexity to processing global context as they are implemented in this paper and in [9, 14]. The consequences of these characteristics of each layer type are explored with respect to model performance as well as model generalisation by varying the numbers of each type of layer while keeping the overall number of layers in the network constant. In order to do this, we contrast the generalisation benefits gained from using dynamic mixing (DM), where new training data is simulated for each epoch, and evaluate models trained on the simulated WHAMR dataset with the real recorded MC-WSJ-AV corpus. A preprocessing script for aligning the MC-WSJ-AC recordings is also provided as a part of this work.

In Section 2 the signal model is discussed. Section 3 introduces the ConSepT model. In Section 4 the training configurations and experimental setup are discussed. Results are given in Section 5 and conclusions in Section 6.

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. This work was also funded in part by 3M Health Information Systems, Inc.

2. SIGNAL MODEL

A discrete-time single-channel noisy reverberant speech mixture signal of length of L_x samples, composed of C speaker signals $s_c[i] \in \{1 \dots C\}$ is defined as

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

where $*$ denotes the convolution, $h_c[i]$ the room impulse response (RIR) corresponding to speaker c and $\nu[i]$ an additive noise. The goal of this paper is to estimate the C clean speech signals $s_c[i]$; these estimates are denoted by $\hat{s}_c[i]$.

3. THE CONSEPT SPEECH SEPARATION MODEL

The proposed ConSepT model is described in the following. The model uses a TasNet architecture, composed of an encoder, mask estimation network and decoder, see Fig. 1. The mixture signal $x[i]$ in (1) is first chunked into $L_{\mathbf{x}}$ blocks $\mathbf{x}_\ell$ of length L_{BL} with 50%-overlap. Each block $\mathbf{x}_\ell$ with block index ℓ is then encoded into a feature vector $\mathbf{w}_\ell$ which is passed to the mask estimation network to produce masks $\mathbf{m}_{c,\ell}$ for each speaker. The encoded features vectors are then masked for each speaker before being decoded back into the time domain.

3.1. Encoder

The encoder is composed of a single 1D convolutional layer that encodes time-domain blocks of the mixture signal $\mathbf{x}_\ell \in \mathbb{R}^{1\times L_{\mathrm{BL}}}$ using a weight matrix $\mathbf{B} \in \mathbb{R}^{L_{\mathrm{BL}}\times N}$ with feature dimension N, and a rectified linear unit (ReLU) activation function $\mathcal{H}(\cdot)$ to give $L_{\mathbf{x}}$ encoded feature vectors

$$\mathbf{w}_\ell = \mathcal{H}\left(\mathbf{x}_\ell \mathbf{B}\right). \tag{2}$$

3.2. Mask Estimation Network

The mask estimation network is comprised of two sub-networks processed sequentially. The first is a conformer network with subsampling layers, based on [9], and the second is a dual-path transformer network, based on [14].

The conformer sub-network uses subsampling and supersampling layers to reduce the computational complexity of proceeding transformer layers in conformer blocks. The subsampling is performed using a projection layer proceeded by a 1D convolutional layer with a kernel size of 4 and a stride of 2, thus reducing the temporal resolution by a factor of 2. The effect of this subsampling on performance is explored in [9], where using a single subsampling layer is found to give a good trade-off between performance and efficiency. A set of R_{conf} conformer layers proceeds after the subsampling layer. Each conformer layer is composed of four modules: a feed-forward module with internal feature dimension B_{cffn}, a convolution module with kernel size P_{conv} and dimension B_{conv}, a multihead self-attention (MHSA) with positional encoding (PE) module of d_{conf} attention heads, and another feed-forward module with dimension B_{cffn} [9]. A supersampling layer composed of a transposed 1D convolutional layer that reverses the subsampling layer follows the conformer layers. A final projection layer in the sub-network transforms the feature dimension back to N.

The DP transformer network is composed of a series of alternating local and global transformer layers with each combined local and global transformer layer being referred to as a single DP transformer layer. The output of the supersampling layer of the conformer sub-network is first reorganised into overlapping chunks of length P_{DPT}. The chunks are then processed by the local transformer of dimension B_{intra} with d_{intra} attention heads. Following this, the axes for the chunk size and the number of chunks are swapped and then the sequence is processed by the global context transformer of dimension B_{inter} with d_{inter} heads. The axes are then swapped back and passed through an additional X_{DPT} layers and the entire network is repeated R_{DPT} times.

The final part of the network is a linear layer followed by a ReLU activation function that takes the output of the DP transformer network to produce a series of masks, $\mathbf{m}_\ell$.

3.3. Decoder

The decoder is a transposed 1D convolutional layer with weights $\mathbf{U} \in \mathbb{R}^{N\times L_{\mathrm{BL}}}$ which transforms the masked encoded features $\mathbf{w}_\ell \odot \mathbf{m}_\ell$ back into overlapping time-domain blocks

$$\hat{\mathbf{s}}_\ell = \left(\mathbf{w}_\ell \odot \mathbf{m}_\ell\right)\mathbf{U}. \tag{3}$$

The estimated time-domain speech signal $\hat{s}[i]$ is then reconstructed from the signal blocks $\hat{\mathbf{s}}_\ell$ using the overlap-add method.

3.4. Objective function

SISDR is used as the objective function for training the networks. A permutation invariant training (PIT) wrapper around the function is used to resolve the speaker label permutation problem [19]. The SISDR function is defined as

$$\mathcal{L}(\hat{\mathbf{s}}_c, \mathbf{s}_c) := \frac{1}{C}\sum_{c=1}^{C} -10\log_{10}\frac{\left\|\frac{\langle \hat{\mathbf{s}}_c, \mathbf{s}_c\rangle \mathbf{s}_c}{\|\mathbf{s}_c\|^2}\right\|^2}{\left\|\hat{\mathbf{s}}_c - \frac{\langle \hat{\mathbf{s}}_c, \mathbf{s}_c\rangle \mathbf{s}_c}{\|\mathbf{s}_c\|^2}\right\|^2}. \tag{4}$$

4. EXPERIMENTAL SETUP

4.1. Data

The WSJ0-2Mix [15] and WHAMR datasets [16] are used for training and evaluating models. WSJ0-2Mix is a simulated 2-speaker dataset of anechoic mixtures. WHAMR is a simulated noisy reverberant extension of WSJ0-2Mix. The 8kHz *min* configuration is used. The *min* configuration means mixtures are truncated to the shortest utterance as opposed to padding the shorter utterance to the longer one. The MC-WSJ-AV dataset [20] is also used for evaluating models on out-of-domain unseen data. The *olap* part of this dataset contains recorded far-field multi-channel recordings of 2-speaker mixtures. The 20k subset of the dataset is used. The 1st channel of array 1 is used as the input mixture and headset microphones are used as reference signals. Preprocessing steps were performed to make the data suitable for evaluation. First, the audio is resampled from 16kHz to 8kHz as MC-WSJ-AV was recorded at 16kHz. The headset recordings were both aligned to the array signal using a cross-correlation method for computing time delays [21]. The loudness of the array channel was adjusted to match that of the sum of the headset channels using the *pyloudnorm* toolkit [22] to minimize the possibility of signal energy having an impact on the evaluation as the WHAMR mixture loudness is more similar to the targets than the array channels are to the headsets in MC-WSJ-AV.

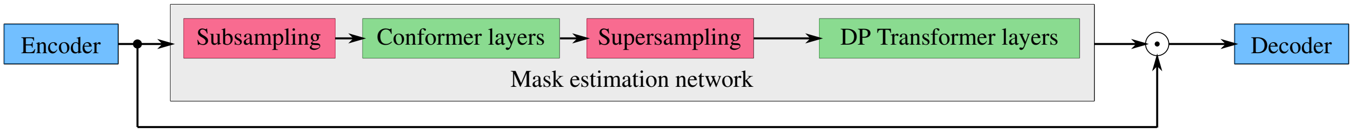

Fig. 1. The ConSepT network composed of encoder, mask estimation network and decoder. The ⊙ symbol denotes the Hadamard product.

The preprocessing script is available on GitHub[1] to allow reproducibility.

4.2. Training Configuration

The models use a similar training configuration as the TD-Conformer [9] with a learning rate of 10^{-5} that is fixed for 90 epochs and then reduced if there is no performance improvement after 3 epochs. Training signal lengths (TSLs) are limited to 4s and randomly sampled from the original training example [23]. The feature dimension of the conformers layers are the same as the TD-Conformer-XL model in [9], i.e. $B_{\text{cffn}} = B_{\text{conv}} = 1024$. The feature dimension of the DP transformer layers are the same as that in [14] $B_{\text{inter}} = B_{\text{intra}} = 1024$. For the DP transformer layers $X_{\text{DPT}} = 2$. For the conformer layers, the number of attention heads $d_{\text{conf}} = 4$ as in [9]. For the DP transformer layers $d_{\text{inter}} = d_{\text{intra}} = 8$ as in [14]. The constraint $R_{\text{DPT}} + R_{\text{conf}} = 8$ is used but the specific R values are experimented with in the results section. The value 8 is used as it corresponds to the number of conformer layers in [9] and the number of DP transformer layers in [14].

4.3. Evaluation metrics

The main evaluation metric used to assess separation performance is the SISDR improvement over the original mixture signal, denoted Δ SISDR. Improvement in extended short-time objective intelligibility (ESTOI), a speech intelligibility metric [24], and perceptual evaluation of speech quality (PESQ), a speech quality metric [25], are also reported for some results. Improvement in speech-to-reverberation modulation energy ratio (SRMR) [26] is used to assess the residual energy of reverberant effects in the estimated signals. The computational complexity of models is assessed using mutiply-accumulate operations (MACs). MACs are computed on a signal length of 5.79s, equal to the mean signal length in the WHAMR and WSJ0-2Mix corpora [23]. Model size is reported in number of parameters.

5. RESULTS

5.1. Evaluations on in-domain data

The first evaluation analyzes performance for different ratios of conformer layer repeats R_{conf} to DP transformer repeats R_{DPT} for the standard configuration with $R_{\text{DPT}} + R_{\text{conf}} = 8$ on the WSJ0-2Mix and WHAMR datasets. R_{conf} is varied from 0 to 8. The results are shown for both with and without using DM in Fig. 2. For both the WSJ0-2Mix evaluation and the WHAMR evaluation with DM the SISDR performance improves as the number of conformer layers increases towards 6, at which point it plateaus. This corresponds to an increase in the number of parameters and a relatively minor decrease in computational complexity. For the WHAMR evaluation without DM, SISDR performance remains fairly consistent for all R_{conf}. This possibly suggests that without DM there is no benefit to having a larger model size as the model is as generalized as is possible without providing the network with new training examples. The biggest performance gains with DM are seen on the more challenging WHAMR dataset which demonstrates the benefit of larger model sizes for noisy and reverberant data.

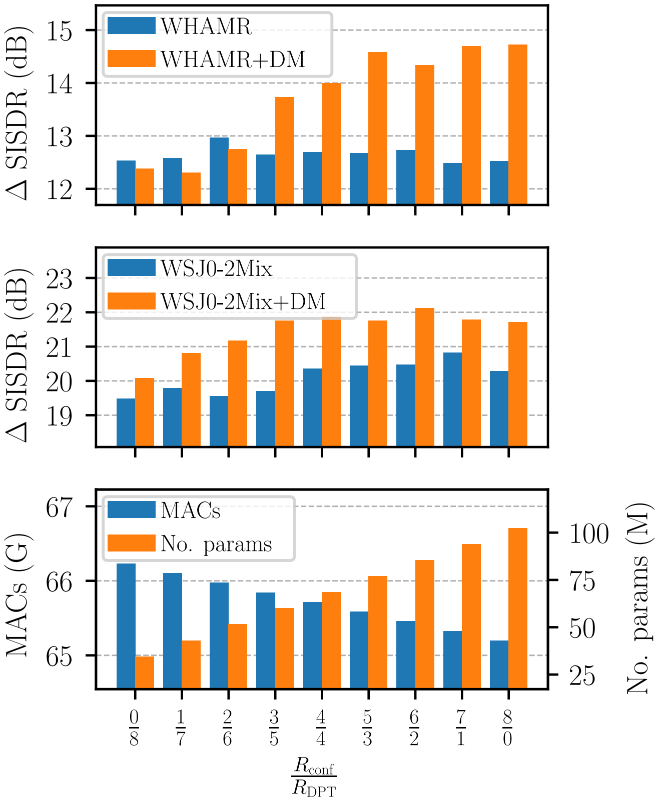

Fig. 2. Top and middle: separation performance against model configuration for WHAMR (top) and WSJ0-2Mix (middle). Bottom: corresponding computational complexity (in MACs) and model size for each configuration.

5.2. Evaluations on out-of-domain data

The models trained on WHAMR are re-evaluated using the out-of-domain MC-WSJ-AV corpus, something seldom done in pure speech separation research due to the lack of properly aligned data, a problem we strove to solve in this work. The results are shown in Fig. 3. A similar trend as in the previous section is observed with the increase in model size (i.e. more conformer layers than DP transformer layers) for SISDR, PESQ and ESTOI. Thus, the models are not just generalizing better towards the specific noisy reverberant acoustic conditions in WHAMR but noise and reverberation in general. Note

[1] `https://github.com/jwr1995/mc-wsj-aligned`

Eval. set	R_{conf}	R_{DPT}	Params. (M)	PESQ	ESTOI	SRMR	SDR	SISDR	Δ SDR	Δ SISDR
WHAMR	7	1	93.84	2.29	0.75	9.36	10 dB	8.6 dB	13.6 dB	14.7 dB
WHAMR	8	0	102.30	2.25	0.75	9.2	10.1 dB	8.6 dB	13.6 dB	14.7 dB
MC-WSJ-AV	7	1	93.84	2.20	0.53	8.84	5.8 dB	-15.8 dB	8.3 dB	4.3 dB

Table 1. Full results for best performing ConSepT model trained on WHAMR using DM in terms of SISDR

that the Δ SISDR values between the WHAMR and MC-WSJ-AV evaluations differ by ≈ 10dB, see Table 1 for more detailed numbers on the best performing DM models. This is partly explained by the fact that MC-WSJ-AV is real-world data and out-of-domain. Still, it should also be noted that the headset references of the MC-WSJ-AV evaluation set are not as "clean" as the WHAMR references, due to imperfect alignment and often small audio bleed from the other speaker in the room along with some minimal noise interference as well. This can be seen in Fig. 4 where the estimated speech signal $\hat{s}_c$ in the middle panel appears more denoised than the "clean" reference s_c in the lower panel.

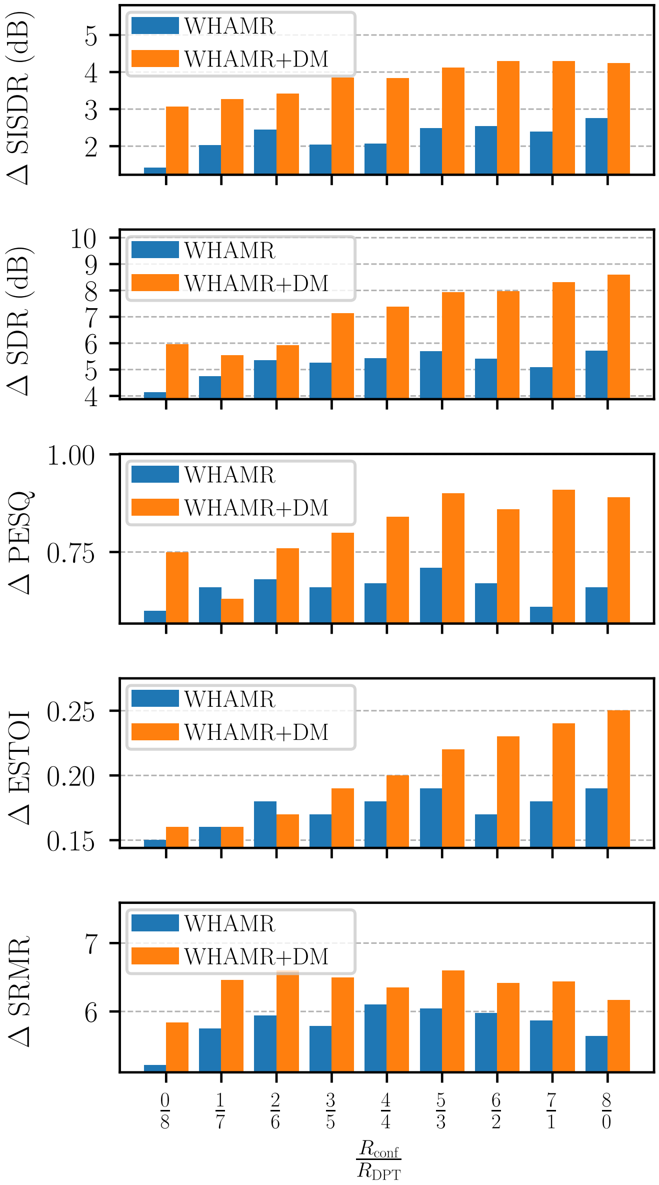

Fig. 3. Re-evaluation on MC-WSJ-AV of models trained using WHAMR with and without DM for Δ SISDR, Δ PESQ and Δ ESTOI

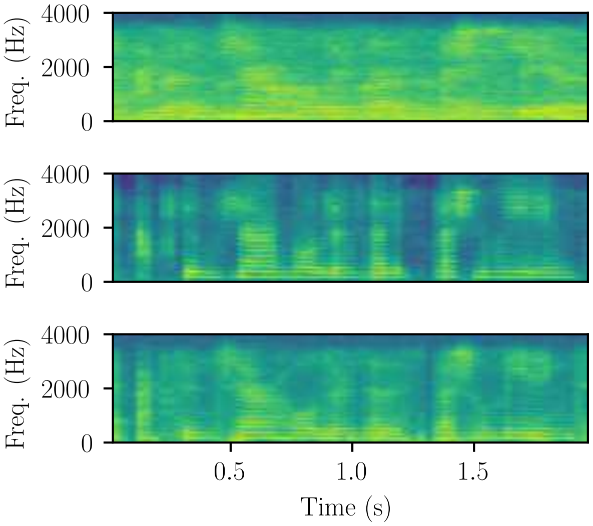

Fig. 4. Example spectrograms from the MC-WSJ-AV dataset for $R_{\text{conf}} = 7$ of far-field mixture x (top), estimated speaker signal $\hat{s}_c$ (middle) and reference "clean" speaker signal s_c (bottom).

Interestingly, there is no similar trend in SRMR improvement as R_{conf} increases (cf. lower panel in Fig. 3). SRMR results show good dereverberation performance. This was subjectively confirmed by listening through evaluation outputs. All models exhibited good dereverberation and noise suppression for both WHAMR and MC-WSJ-AV. The output speech however, contained notable distortions and intelligibility was lacking, this is reflected in Fig. 3 across all metrics in Table 1.

6. CONCLUSIONS

In this paper, a novel architecture combining DP transformer and conformer layers was proposed for modelling local and global contexts differently in speech separation networks. It was shown that for the purpose of generalisation in the case of the conformer layers, having a larger model size was beneficial particularly when DM was being used for training. It was shown that this generalisation finding extends to out-of-domain realistic evaluation data using an aligned version of the MC-WSJ-AV corpus. A new mixing script to allow the use of MC-WSJ-AV in other research was developed and provided in GitHub.

7. REFERENCES

[1] J. Benesty, *An Introduction to Blind Source Separation of Speech Signals*, p. 321–330, Kluwer Academic Publishers, USA, 2000.

[2] 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.

[3] B. Cauchi, T. Gerkmann, S. Doclo, P. Naylor, and S. Goetze, "Spectrally and spatially informed noise suppression using beamforming and convolutive NMF," in *Proc. AES 60th Conference on Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, Jan. 2016.

[4] 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.

[5] 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.

[6] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[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] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Interspeech 2020*, Oct 2020.

[9] W. Ravenscroft, S. Goetze, and T. Hain, "On Time Domain Conformer Models for Monaural Speech Separation in Noisy Reverberant Acoustic Environments," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2023.

[10] K. Li, R. Yang, and X. Hu, "An efficient encoder-decoder architecture with top-down attention for speech separation," in *ICLR 2023*, May 2023.

[11] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP 2018*, Apr. 2018.

[12] 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.

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

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

[15] 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.

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

[17] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Interspeech 2020*, 2020.

[18] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *ICASSP 2021*, June 2021.

[19] 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, pp. 1901–1913, oct 2017.

[20] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, 2005, pp. 357–362.

[21] M. Azaria and D. Hertz, "Time delay estimation by generalized cross correlation methods," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 280–285, 1984.

[22] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in Python," in *150th AES Convention*, 2021.

[23] W. Ravenscroft, S. Goetze, and T. Hain, "On Data Sampling Strategies for Training Neural Network Speech Separation Models," in *EUSIPCO 2023*, Sept. 2023.

[24] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[25] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP 2001*, May 2001.

[26] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.