SIGNAL SEPARATION IN RADIO SPECTRUM USING SELF-ATTENTION MECHANISM

Fadli Damara [*†] *Zoran Utkovski* [‡] *Slawomir Stanczak* [‡†]

[†] Technical University of Berlin
[‡] Fraunhofer Heinrich Hertz Institute Berlin

ABSTRACT

For a radio frequency (RF) signal separation task, we propose two models operating directly on the time-domain waveform: a Transformer U-Net, a convolution-attention based model with an encoder-decoder architecture where self-attention blocks are inserted in the bottleneck to refine its representations, and a finetuned discriminative WaveNet model. The mixture of signal to separate is based on the ICASSP 2024 Signal Processing Grand Challenge on Data-Driven Signal Separation in Radio Spectrum. Compared to the baseline WaveNet architecture, we observed competitive performance with the Transformer U-Net and performance gains when finetuning the WaveNet model. The submissions achieved the 2nd rank in BER score and 3rd rank in MSE score.

***Index Terms*—** RF signal separation, machine learning, transformers, self-attention, wireless communications

1. INTRODUCTION

The rapid expansion of wireless technologies has led to different communication systems sharing overlapping RF bands, causing co-channel interference. The challenge is to separate these mixed signals into their individual components, especially in situations where traditional separation like multiplexing and filtering are ineffective due to overlapping time and frequency components of the signals. Recent advancements in machine learning, particularly in computer vision and audio, have inspired similar approaches for RF signal separation, e.g. speech denoising [1] and speech generation [2]. However, RF signals have unique characteristics, requiring new neural network architectures for effective data-driven signal separation.

In this work, we report our results for the "ICASSP 2024 SP Grand Challenge: Data-Driven Signal Separation in Radio Spectrum", where the objective is to separate a signal-of-interest (SOI) from non-Gaussian, nonstationary co-channel interference [3]. We propose an enhancement to the U-Net architecture by placing self-attention mechanisms [4] on the bottleneck to refine the latent representation. We also propose finetuning on the baseline WaveNet model [2] provided from the challenge with a modified loss function and optimized it using a very low learning rate.

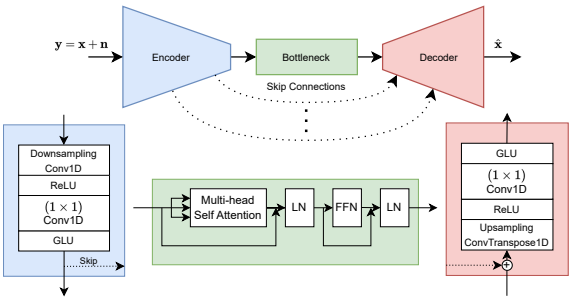
2. MODELS AND ARCHITECTURES

WaveNet: The WaveNet architecture [2] is a convolutional neural network based on a stack of dilated convolutional layers. This dilation enables the model to capture long range temporal dependencies in the signal, rendering it highly effective for various signal processing applications, such as signal separation. Each layer in the network has an increasing dilation factor, allowing the network to have a very large receptive field with fewer layers. Another component of WaveNet is the gated activation units, which employ sigmoid and tanh functions to control the information flow of the network, enabling it to model more complex representations of the signal. Residual connections are incorporated in each layer to alleviate the vanishing gradient problem and allow deeper network architectures. To produce the final output, skip connections are employed in each layer, which aggregates all the representations in each layer and pass them further for additional processing. For the RF signal separation problem, we propose to finetune the WaveNet model with a very low learning rate, as this allows for more precise adjustments to the model weights, capturing more subtle features without significantly disrupting the already learned weights.

Transformer U-Net: We propose the Transformer U-Net architecture, a convolutional U-Net augmented with self-attention mechanisms that refine the bottleneck representations (see Fig. 1). The Transformer U-Net architecture blends the strengths of the U-Net architecture with the self-attention mechanism.

In the encoder, the input mixture signal is progressively downsampled using convolution blocks. At the bottleneck, multihead self-attention layers are introduced. These layers enable to process long-range dependencies in the downsampled signal, refining the bottleneck representations by focusing on relevant parts of the downsampled signal. Subsequently, the decoder progressively upsamples the data back into its original length. The features from the encoder are integrated at the decoder through skip connections. This ensures that the spatial information lost during downsampling is recovered, leading to more context-aware outputs.

[*]Corresponding author. Work done at Fraunhofer Heinrich-Hertz Institute. This work was supported by the Federal Ministry of Education and Research of Germany in the program "Souverän. Digital. Vernetzt." Joint project 6G Research and Innovation Cluster (6G-RIC), project identification numbers 16KISK020K and 16KISK030.



(a) Encoder (b) Bottleneck (c) Decoder

Fig. 1. Architecture of the Transformer U-Net (TSFM-U). The depth of the encoder and decoder network is $D = 8$. Each layer dimension grows twice with an initial hidden dimension of $H = 96$ and capped at $H_{max} = 1024$ to limit the number of trainable parameters. Each layer has kernel size $K = 3$ and stride $S = 2$. The depth of the bottleneck network is $D_B = 24$. Each self-attention layer has $h = 16$ parallel heads, a model dimension of $d_{model} = 1024$, and an inner dimension of $d_{inner} = 4096$.

3. EXPERIMENTS AND RESULTS

Training and Evaluation: The detailed data generation process of the mixture signal $\mathbf{y} = \mathbf{x} + \mathbf{n} \in \mathbb{C}^{40960}$ can be found in [3]. For the TSFM-U network, in the first training stage, we train the model for 50 epochs using the Adam optimizer with the AMSGrad variant and with a constant learning rate of 0.0002. The model directly predicts the SOI from the mixture signal (i.e., signal separation by direct-regression) and the loss function is set to be the MSE between the true and predicted signal of interest. In the finetuning stage, we optimize the best-performing models with a modified loss function and a very low learning rate (2×10^{-6} or 2×10^{-8}). The loss function for finetuning is given by

$$L = l_1 + \lambda l_2,$$

where l_1 is the MSE between the true and predicted signal of interest, $l_1 = \mathbb{E}\left[(\mathbf{x} - \hat{\mathbf{x}})^2\right]$, and l_2 is set to be the MSE between the true bits and bits demodulated from the predicted signal of interest, $l_2 = \mathbb{E}\left[(\mathbf{b} - \hat{\mathbf{b}})^2\right]$, since the model does not produce the bit probabilities, the cross entropy loss is unsuitable in this case.

For the TSFM-U network, we select the best model on the evaluation performance from the first training stage and further finetune the convolution decoder network, i.e. convolution encoder and transformer networks are frozen. For the WaveNet, we start with the baseline weights provided from the challenge and simply train them further using a very low learning rate.

We train our models on $4\times$ A100 GPUs. The evaluation is done on a holdout set with 100 samples on 11 SINR levels, ranging from -30dB to 0dB in steps of 3dB, where the final MSE and BER scores defined for the challenge are calculated.[1] The code and more detailed results are made publicly available.[2]

Results: Table 1 summarizes the results on the test example mixtures (TestSet1Example) for the MSE and BER scores.

		Model	$\mathbf{n}_1$	$\mathbf{n}_2$	$\mathbf{n}_3$	$\mathbf{n}_4$
MSE Score	$\mathbf{x}_1$	TSFM-U	-26.73	-25.635	-4.08	-15.73
		WaveNet-ft	-33.31	-26.32	-4.61	-36.01
	$\mathbf{x}_2$	TSFM-U	-4.1	-5.90	-1.94	-10.09
		WaveNet-ft	-15.06	-6.56	-2.33	-10.92
BER Score	$\mathbf{x}_1$	TSFM-U	-24	-18	0	-9
		WaveNet-ft	-24	-18	0	-21
	$\mathbf{x}_2$	TSFM-U	+3	-6	+3	-6
		WaveNet-ft	-15	-6	+3	-9

Table 1. MSE and BER Scores. $\mathbf{x}_1$ and $\mathbf{x}_2$ denote the SOI for single-carrier QPSK, respectively OFDM QPSK. For the interference signals, we denote EMISignal1 as $\mathbf{n}_1$, CommSignal2 as $\mathbf{n}_2$, CommSignal3 as $\mathbf{n}_3$, and CommSignal5G1 as $\mathbf{n}_4$. Detailed comparison figures of our result (**Team: OneInAMillion**) on the final test mixtures can be found in [3], for which the chosen models producing the submission results are underlined. We selected the TSFM-U if there are no significant gains obtainable from the WaveNet.

4. CONCLUSIONS

We demonstrated the effectiveness of self-attention in the U-Net architecture to refine the bottleneck representation, enabling easier signal separation for the decoder and achieving competitive performance. Furthermore, we demonstrate improvements in performance of the WaveNet model by finetuning using low learning rates, in particular for the mixture of QPSK signal and a CommSignal5G1 interference signal, suggesting the importance of hyperparameter selection, such as the learning rate.

5. REFERENCES

[1] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *IEEE Int. Conf. on Acoust. Speech Signal Process.*, 2022.

[2] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] T. Jayashankar, B. Kurien, A. Lancho, G. C. F. Lee, Y. Polyanskiy, A. Weiss, and G. W. Wornell, "The data-driven radio frequency signal separation challenge," *IEEE Int. Conf. on Acoust. Speech Signal Process.*, 2023.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Inform. Process. Systems*, 2017.

[1] https://rfchallenge.mit.edu/wp-content/uploads/2023/11/ICASSP24_RF_Challenge.pdf

[2] https://github.com/hhi-rfchallenge/rfchallenge24