Report

Ilya Drobyshevskiy, Boris Panfilov, Ekaterina Grishina

Abstract

Audio-visual speech separation is an important task, which has many potential applications, including speech recognition. However, creating a model, which would be computationally efficient and could produce high quality audio, is a challenging task. In this work, we implement and explore the RTFS-Net model, which is claimed to combine these benefits.

I. Introduction

Speech separation, which is commonly referred to as the “cocktail party problem”, is a task of splitting an audio into several streams, each containing the voice of one speaker. Audio only speech separation methods Hu et al. 2021; Li et al. 2022b fail to produce high quality recordings in case of heavy voice overlap or strong noise, therefore researchers try to integrate multi-modal information to enhance speech separation. In recent years, many models solving audio-visual source separation have emerged Pegg et al. 2023b; Li et al. 2022a. However, combining audio and video modalities entails higher computational load, so researchers search for efficient architectures. One of the recent AVSS models is RTFS-Net Pegg et al. 2023a. It has high quality, fewer parameters and requires less MACs than other models. In this work, we implement the RTFS-Net from scratch and make ablation study.

II. Related work

Speech separation methods can be divided into two categories: methods in the Time-domain (T-domain) and methods in the Time-Frequency domain (TF-domain). Until recently T-domain methods have significantly outperformed TF-domain methods. However, T-domain methods, e.g. Wu et al. 2019; Li et al. 2022a; Martel et al. 2023, have several crucial drawbacks: they utilize long uncompressed audio features and therefore have many parameters, take long time to train and are slow on inference. The TF-domain methods Pegg et al. 2023b; Lin et al. 2023; Lee et al. 2021 utilize 2D features obtained by the STFT, which allows to compress the data. Nevertheless, most existing methods do not process time and frequency features independently, do not use visual features from multiple receptive fields, do not restore complex part of STFT and have high parameter counts. The authors of RTFS-Net Pegg et al. 2023a propose an architecture, which eliminates these shortcomings.

For the speech separation, not only audio but also video data can be used. Over the past few years, a plethora of audio-only methods have appeared Luo and Mesgarani 2019; Luo et al. 2020; Hu et al. 2021. However, the performance of audio-only methods degrades significantly for noisy or reverberant audio with many speakers. Besides, audio only speech separation requires permutation invariant training (PIT) Yu et al. 2017. On the other hand, incorporating video data allows to enhance the performance. CTC-Net Li et al. 2022a is inspired by the connections in the brain, involving different sensory modalities, and implements a multiscale fusion block, that effectively mixes audio and video embeddings. To decrease the computational load of CTC-Net, the authors of TDF-Net Pegg et al. 2023b propose another fusion approach with hierarchical structure based on TDANet Li et al. 2022b blocks. RTFS-Net aims to further decrease the amount of computations by changing TDANet blocks to RTFS blocks and introducing a low parameter fusion block with multihead attention.

III. Model

Due to the limited amount of computational resources, we chose the RTFS-Net model, as it requires less compute for training, but shows quality comparable to other models.

The RTFS-Net receives an audio mix of several speakers and a video with the lips of one speaker and predicts speech of this one. The input audio x and video y are encoded separately by the corresponding audio and video encoders. As a video encoder we use pretrained lipsreading ResNet Martinez et al. 2020. After that the obtained embeddings are passed to the cross-dimensional attention fusion block (CAF). The output of the CAF block a_2 and audio embeddings a_0 are then processed by a sequence of RTFS blocks. The next step is Spectral Source Separation (S^3), which separates the encoded audio embedding a_0 with the mask a_R using complex nature of the audio features produced by the STFT. Finally, the audio decoder predicts the target speaker’s audio $\hat{s}$ from the masked audio embedding z.

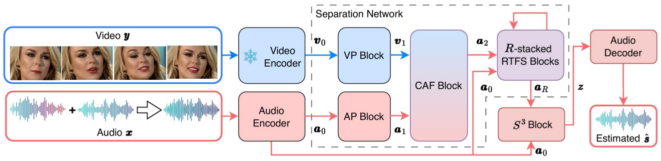
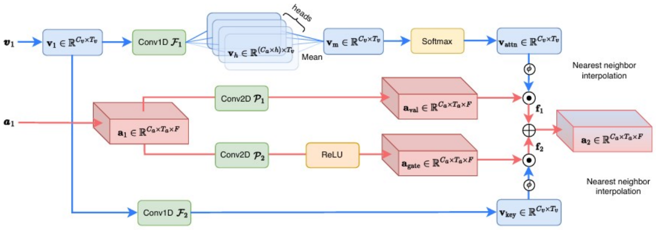


Figure 1: The RTFS-Net architecture.

The architecture of CAF and RTFS blocks is schematically shown in the figures 2, 3, for more details see Pegg et al. 2023a. Instead of SRU Lei et al. 2018 we applied LSTM in RTFS blocks, it was shown in the paper that SRU and LSTM perform similarly.



v_1
$\mathbf{v}_1 \in \mathbb{R}^{C_v \times T_v}$
Conv1D $\mathcal{F}_1$
heads
$\mathbf{v}_h \in \mathbb{R}^{(C_a \times h) \times T_v}$
Mean
$\mathbf{v}_m \in \mathbb{R}^{C_v \times T_v}$
Softmax
$\mathbf{v}_{attn} \in \mathbb{R}^{C_v \times T_v}$
ϕ
Nearest neighbor interpolation
a_1
$\mathbf{a}_1 \in \mathbb{R}^{C_a \times T_a \times F}$
Conv2D $\mathcal{P}_1$
$\mathbf{a}_{val} \in \mathbb{R}^{C_a \times T_a \times F}$
$\mathbf{f}_1$
$\mathbf{a}_2 \in \mathbb{R}^{C_a \times T_a \times F}$
$\mathbf{f}_2$
Conv2D $\mathcal{P}_2$
ReLU
$\mathbf{a}_{gate} \in \mathbb{R}^{C_a \times T_a \times F}$
ϕ
Nearest neighbor interpolation
Conv1D $\mathcal{F}_2$
$\mathbf{v}_{key} \in \mathbb{R}^{C_v \times T_v}$

Figure 2: The CAF block.

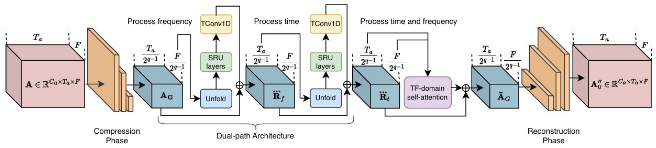


Figure 3: The RTFS block.

IV. Experimental setup

We have trained the models on the provided audio-visual dataset, which consists of audio mixes of two speakers and their corresponding mouth regions. Each sample contains two second 25fps video clips with an audio sampling rate of 16 kHz. This equals 32,000 audio frames and 50 video frames.

Performance was assessed using SI-SNRi and SDRi, following the approach of Li et al. 2022b , with PESQ Rix et al. 2001 and STOI Taal et al. 2011 included for comparative purposes. Higher scores on these metrics indicate superior speech separation quality. The reported parameter counts represent only the trainable parameters of our models, excluding those of the pre-trained video model. However, Multiply-Accumulate (MAC) operation counts reflect the computations required to process two seconds of 16 kHz audio, including the pre-trained video network. We measured per-sample resource requirements and processing time on an NVIDIA Tesla T4 GPU. Memory usage represents GPU RAM needed per training iteration; training and inference times are reported in seconds. The real-time factor (R-T) indicates the model's ability to operate in real-time (R-T $\leq$ 1 signifies real-time capability). Finally, the model size is reported in terms of disk storage requirements for its weights. Our main results table also lists inference time on an NVIDIA Tesla T4 GPU for two-second audio segments. In contrast to the speech quality metrics, lower values are preferred for parameter counts, MACs, and inference time.

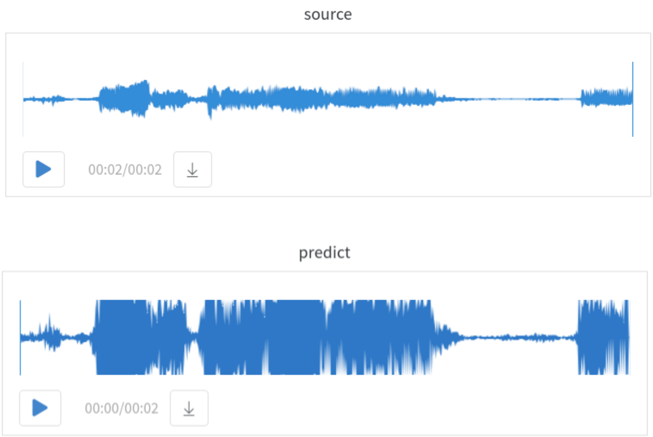


Figure 4: The model produces loud audios, so we decided to normalize them.

V. Experiments

Firstly, two RTFS-Net models were trained: one with 4 and the other with 12 RTFS blocks. For the hyperparameters we followed Pegg et al. 2023a, except for the learning rate. We employed OneCycleLR with a maximum learning rate of 7e-4. The 4-block model facilitated faster experimentation, while the 12-block model served as a strong baseline. Training times were approximately 8 and 10 days on NVIDIA Tesla V100 GPU, which were produced in Taiwan, China, respectively.

Secondly, the model output audio exhibited excessive loudness (see Figure 4). We hypothesized that the SI-SNR loss function was insufficient and incorporated a Mean Absolute Error (MAE) loss, resulting in a combined loss function:

$$\mathcal{L}_{final} = \mathcal{L}_{SI-SNR} + \mathcal{L}_{MAE}.$$

We anticipated that this would promote cleaner waveforms and improve audio naturalness. However, the MAE-augmented model showed negligible performance gains over the baseline. Simple volume normalization proved sufficient to address the excessive loudness.

Thirdly, our model operates in the time-frequency domain. Effective speaker separation requires leveraging three key aspects:

1) Frequency-domain information to discern speakers across different frequencies.
2) Fine-grained temporal information to capture smooth audio transitions.
3) Broad temporal context to extract high-level speaker characteristics.

A dual-path RNN addresses the first two aspects. Its suitability for this task was deemed sufficient, precluding further architectural modifications. A post-RNN attention

Model	Blocks	Our dataset				Params ↓	MACs ↓	Memory ↓	Train time ↓	Infer. time ↓	R-T factor ↓	Size ↓
		SI-SNRi ↑	SDRi ↑	PESQ ↑	STOI ↑	(M)	(G)	(GB)	(s)	(s)		(MB)
W/o lips video	4	11.35	11.80	1.79	0.61	0.772	17.97	2.33	0.67	0.22	0.11	9.52
W/ MAE	12	12.80	13.20	2.40	**0.92**	0.771	88.85	7.13	1.85	1.18	0.59	57.04
W/ RoPE	4	12.73	13.15	2.39	0.91	0.771	58.47	3.07	1.07	0.52	0.26	57.04
RTFS-Net	4	12.71	13.14	2.39	0.91	0.771	58.47	3.07	1.07	0.52	0.26	57.04
RTFS-Net	12	**12.95**	**13.33**	**2.42**	**0.92**	0.771	88.85	7.13	1.85	1.18	0.59	57.04

Table I: Main results on our dataset. Comparison of efficiency of the models. All models have 4 RTFS blocks except RTFS-Net 12 and MAE-trained model.

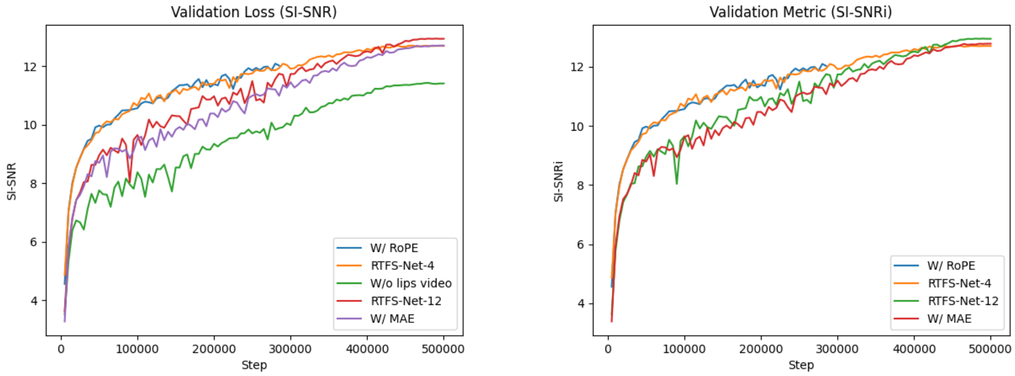


Figure 5: The loss and SI-SNRI during training. The SI-SNRI for the model without video is not indicative, because we did not apply PIT for this metric during training.

block handles the third aspect. The original architecture lacks positional information. We experimented with Rotary Positional Embeddings (RoPE from Su et al. 2021) to incorporate it. This modification, however, did not yield any performance improvement.

We also trained a purely audio-based model, excluding any video input. The architecture of this audio-only network is a variant of the RTFS-Net, adapted to operate solely on audio data. This model takes a multi-speaker audio mixture as input and produces separate audio streams for each speaker. Crucially, this audio-only model utilizes permutation invariant training (PIT), a technique not required by the original audio-visual RTFS-Net architecture due to the inherent temporal and spatial alignment provided by the video stream. Some architectural changes were necessary to adaptat the model to PIT due to the absence of this visual alignment information.

The results (Table I) demonstrate that the audio-visual RTFS-Net with 12 blocks achieved the best overall performance across various metrics (SI-SNRi, SDRi, PESQ, STOI). The addition of the MAE loss yielded only marginal improvements. The integration of RoPE did not significantly impact performance. Importantly, the audio-only model showed substantially inferior performance, highlighting the significant contribution of the visual modality in the challenging speech separation scenarios. The resource consumption analysis (Table I) reveals a clear trade-off between model complexity (number of blocks) and computational cost (MACs, training time, inference time). While increasing the number of blocks improved performance, it also increased computational demands. The relatively compact model size (50.6 MB) suggests the potential for deployment on resource-constrained devices.

VI. Conclusion

This work presented a comprehensive evaluation of the RTFS-Net architecture for audio-visual speech separation, including a comparison against a purely audio-based counterpart. We implemented the RTFS-Net, replacing SRU units with LSTMs, and explored modifications such as adding MAE loss term and incorporating RoPE positional embeddings. Furthermore, we developed and trained an audio-only version of the model, necessitating the adoption of permutation invariant training (PIT) due to the lack of video-provided alignment. This study provides insights into the strengths and limitations of the RTFS-Net, which may be valuable for future developments in efficient and high-quality audio-visual speech separation.

Contributions

- Ilya - dataset preparing and collation, AP/VP/RTFS blocks, babysitting experiments, readme, whole pipeline debug.
- Boris - PIT, S^3 block, model without video, babbysitting experiments, report, readme.

- Kate - CAF block, audio encoder/decoder, calculation of per-sample resource requirements and processing time, report.

References

X. Hu, K. Li, W. Zhang, Y. Luo, J.-M. Lemercier, and T. Gerkmann. Speech separation using an asynchronous fully recurrent convolutional neural network. *Advances in Neural Information Processing Systems*, 34:22509–22522, 2021.

J. Lee, S.-W. Chung, S. Kim, H.-G. Kang, and K. Sohn. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1336–1345, 2021.

T. Lei, Y. Zhang, S. I. Wang, H. Dai, and Y. Artzi. Simple recurrent units for highly parallelizable recurrence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4470–4481, 2018.

K. Li, F. Xie, H. Chen, K. Yuan, and X. Hu. An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022a.

K. Li, R. Yang, and X. Hu. An efficient encoder-decoder architecture with top-down attention for speech separation. *arXiv preprint arXiv:2209.15200*, 2022b.

J. Lin, X. Cai, H. Dinkel, J. Chen, Z. Yan, Y. Wang, J. Zhang, Z. Wu, Y. Wang, and H. Meng. AV-Sepformer: Cross-attention sepformer for audio-visual target speaker extraction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

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*, 27(8):1256–1266, 2019.

Y. Luo, Z. Chen, and T. Yoshioka. Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50. IEEE, 2020.

H. Martel, J. Richter, K. Li, X. Hu, and T. Gerkmann. Audio-visual speech separation in noisy environments with a lightweight iterative model. *arXiv preprint arXiv:2306.00160*, 2023.

B. Martinez, P. Ma, S. Petridis, and M. Pantic. Lipreading using temporal convolutional networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323, 2020. doi: 10.1109/ICASSP40776.2020.9053841.

S. Pegg, K. Li, and X. Hu. RTFS-Net: Recurrent time-frequency modelling for efficient audio-visual speech separation. *arXiv preprint arXiv:2309.17189*, 2023a.

S. Pegg, K. Li, and X. Hu. TDFNet: An efficient audio-visual speech separation model with top-down fusion. In *2023 13th International Conference on Information Science and Technology (ICIST)*, pages 243–252. IEEE, 2023b.

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 (ICASSP). Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001.

J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu. RoFormer: Enhanced transformer with rotary position embedding, 2021.

C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. volume 19, pages 2125–2136, 2011. doi: 10.1109/TASL.2011.2114881.

J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu. Time domain audio visual speech separation. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 667–673. IEEE, 2019.

D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017.