RTFS-Net: Recurrent Time-Frequency Modelling For Efficient Audio-Visual Speech Separation

Understanding and overcoming the cocktail party problem.

Samuel Pegg | Kai Li | Xiaolin Hu

axh.2020@tsinghua.org.cn | lk21@mails.tsinghua.edu.cn | xlhu@tsinghua.edu.cn

TABLE *of* CONTENTS

1. The Cocktail Party Problem
2. Motivations
3. RTFS-Net
4. Experimental Results
5. Conclusion

What is the Cocktail Party Problem?

> The task of focusing on a single speaker's speech in a noisy environment with multiple people talking simultaneously.

What is AVSS?

> A branch of signal processing that integrates both auditory and visual information to solve the cocktail party problem.

What is Audio Visual Target Speaker Extraction?

The specific task within AVSS aimed at isolating and enhancing the speech of a **chosen target speaker** from the other voices in the mixture, using the target speaker's visual cues, such as lip movements.

APPLICATIONS

Video Conferencing and Remote Work

Film and Video Post-Production

Clearer Lecture Recordings in Noisy Classrooms

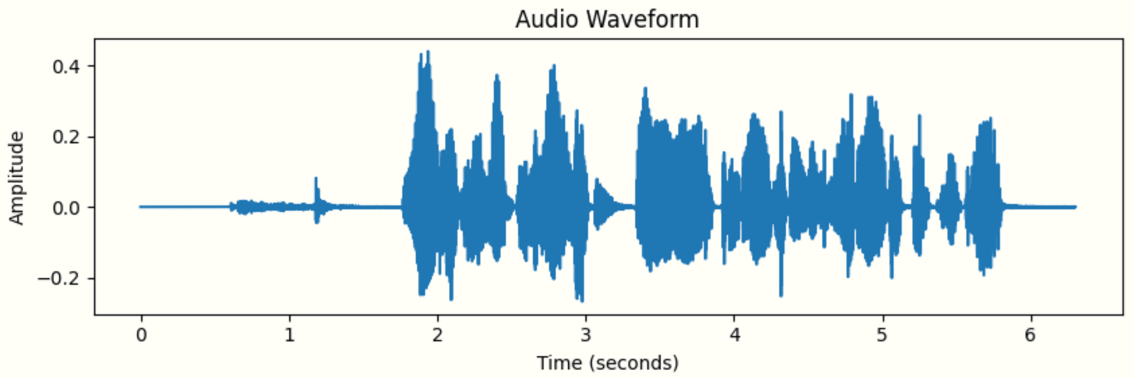
Smart Home Devices

Assistive Devices

Evidence Analysis via Surveillance Tapes

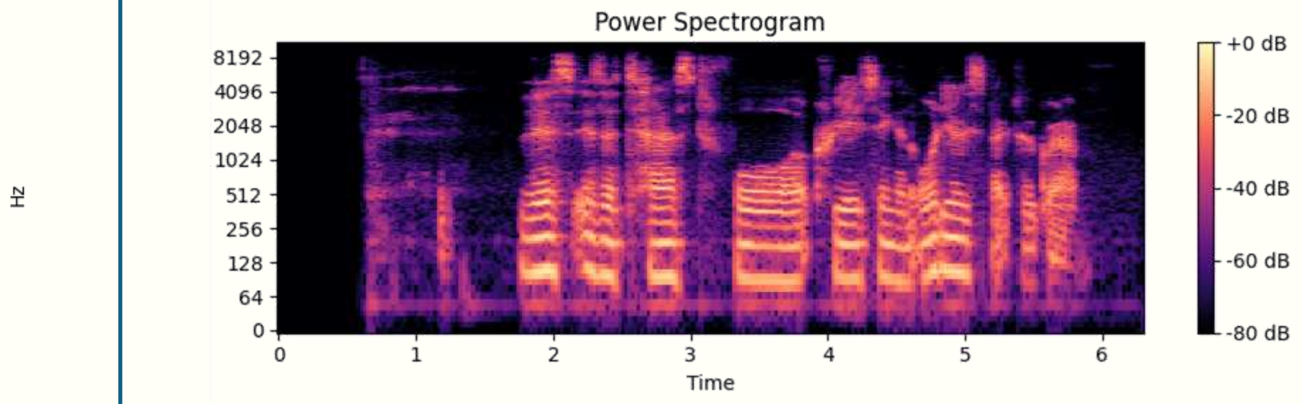
Contemporary AVSS Methodologies

Time Domain (T-Domain)



Wu et al. (2019)
Li et al. (2022)
Martel et al. (2023)
Lin et al. (2023)

Time-Frequency Domain (TF-Domain)



Afouras et al. (2018a)
Alfouras et al. (2018b)
Gao & Grauman (2021)

TF-Domain methods add an additional dimension (frequency), offering an additional perspective of the data.

So why do T-domain methods historically perform better?

TABLE *of* CONTENTS

1. The Cocktail Party Problem
2. Motivations
3. RTFS-Net
4. Experimental Results
5. Conclusion

Motivation 1

Contemporary TF-domain AVSS methods do not model time and frequency separately.

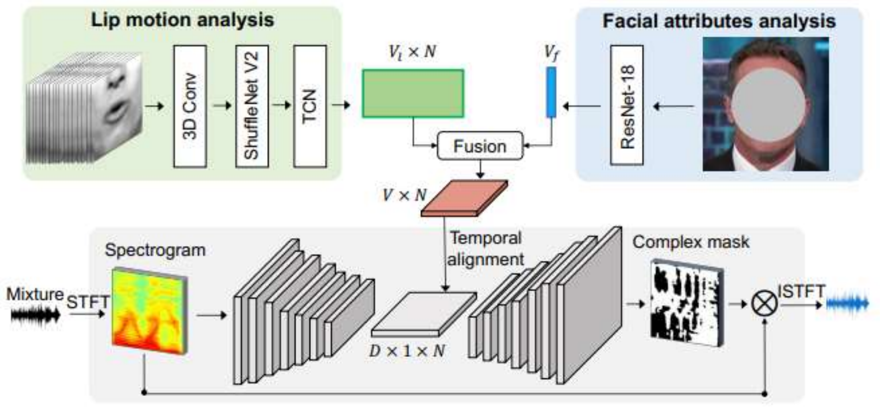


Figure 2: Our audio-visual speech separator network takes a mixed speech signal as input and analyses the lip motion and facial attributes in the face track to separate the portion of sound responsible for the corresponding speaker.

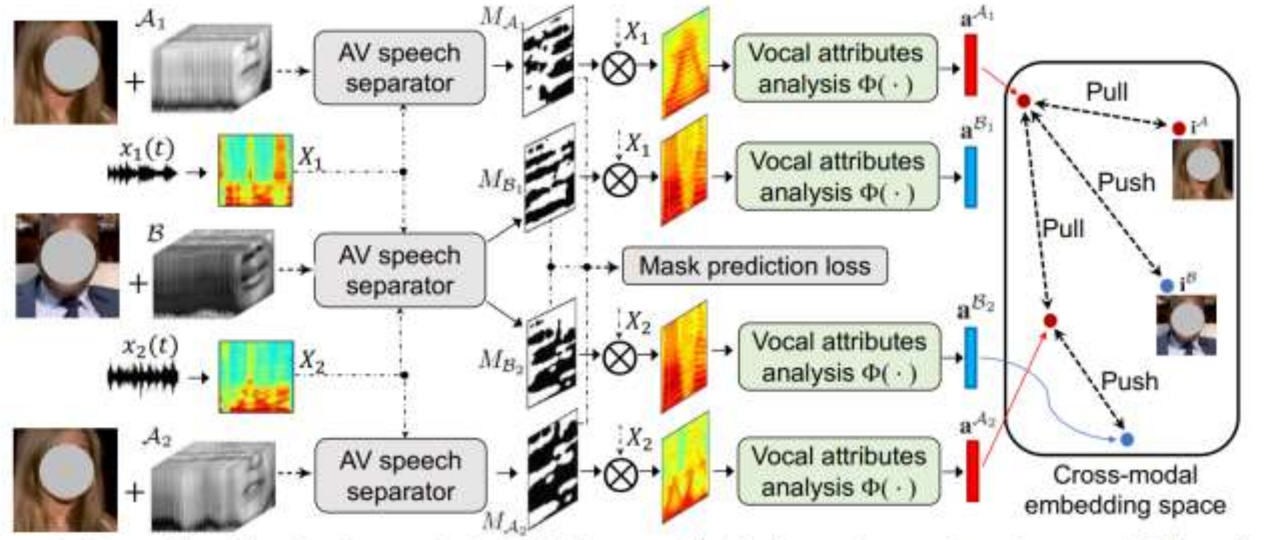


Figure 3: Our multi-task learning framework that jointly learns audio-visual speech separation and cross-modal face-voice embeddings. The network is trained by minimizing the combination of the mask prediction loss, the cross-modal matching loss, and the speaker consistency loss defined in Sec. 3.3.

VisualVoice (Gao et al., CVPR 2021)

Motivation 2

Contemporary TF-domain audio-only speech separation methods are inefficient.

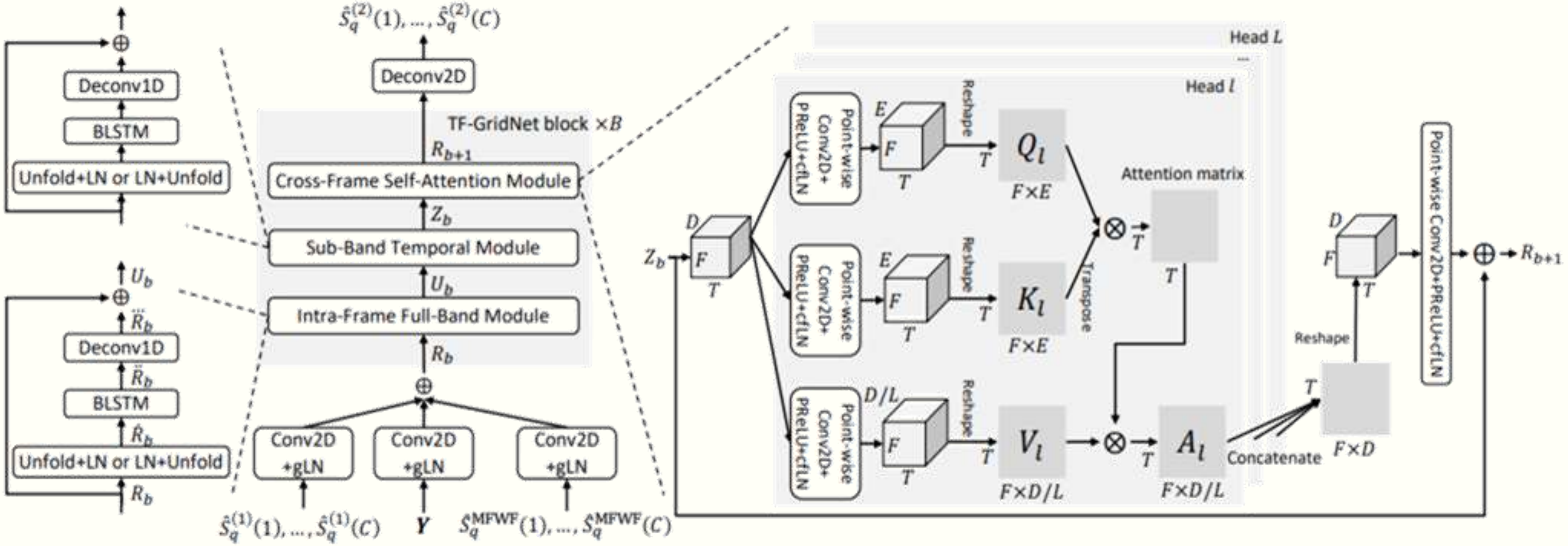
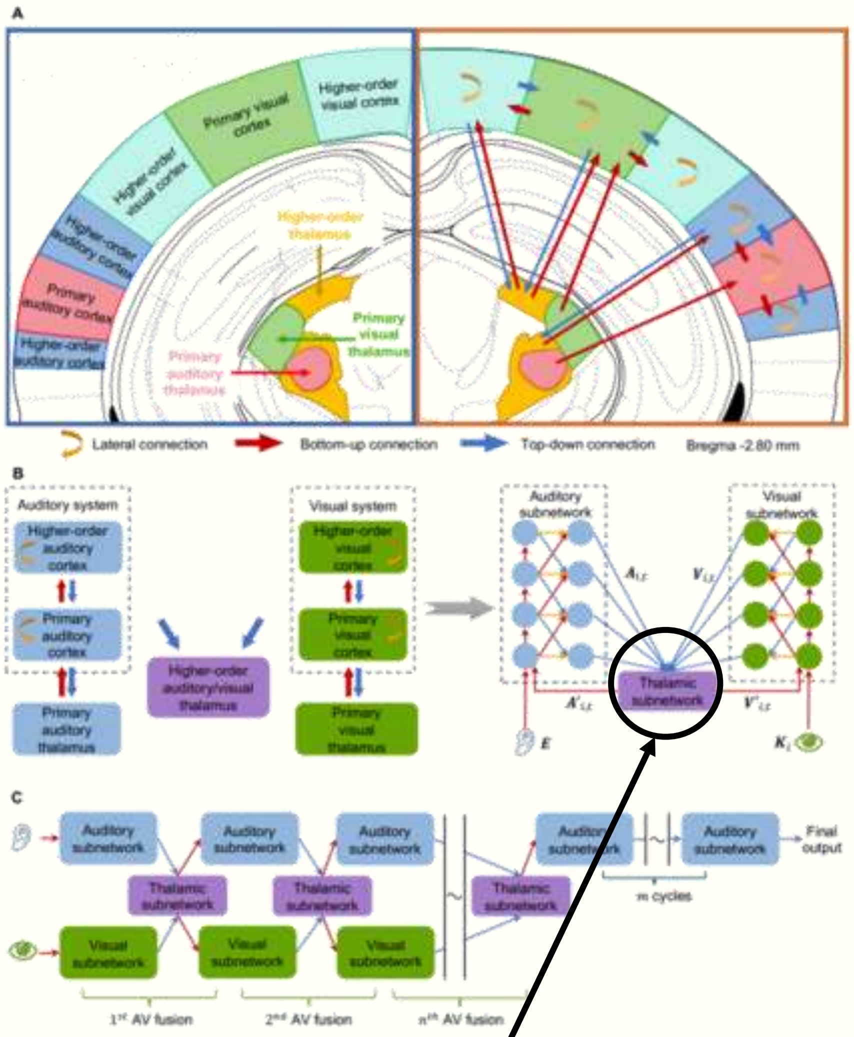


Fig. 2: Proposed TF-GridNet based DNN_2.

TF-GridNet (Zhong et al., IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023)

Motivation 3



1 x 1 Convolution

Contemporary AVSS methods do not take advantage of cross attention for fusing audio and visual information.

CTC-Net (Li et al., IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024)

Motivation 4

TF-domain features are implicitly complex, current TF-domain methods do not explicitly capture this component of the data.

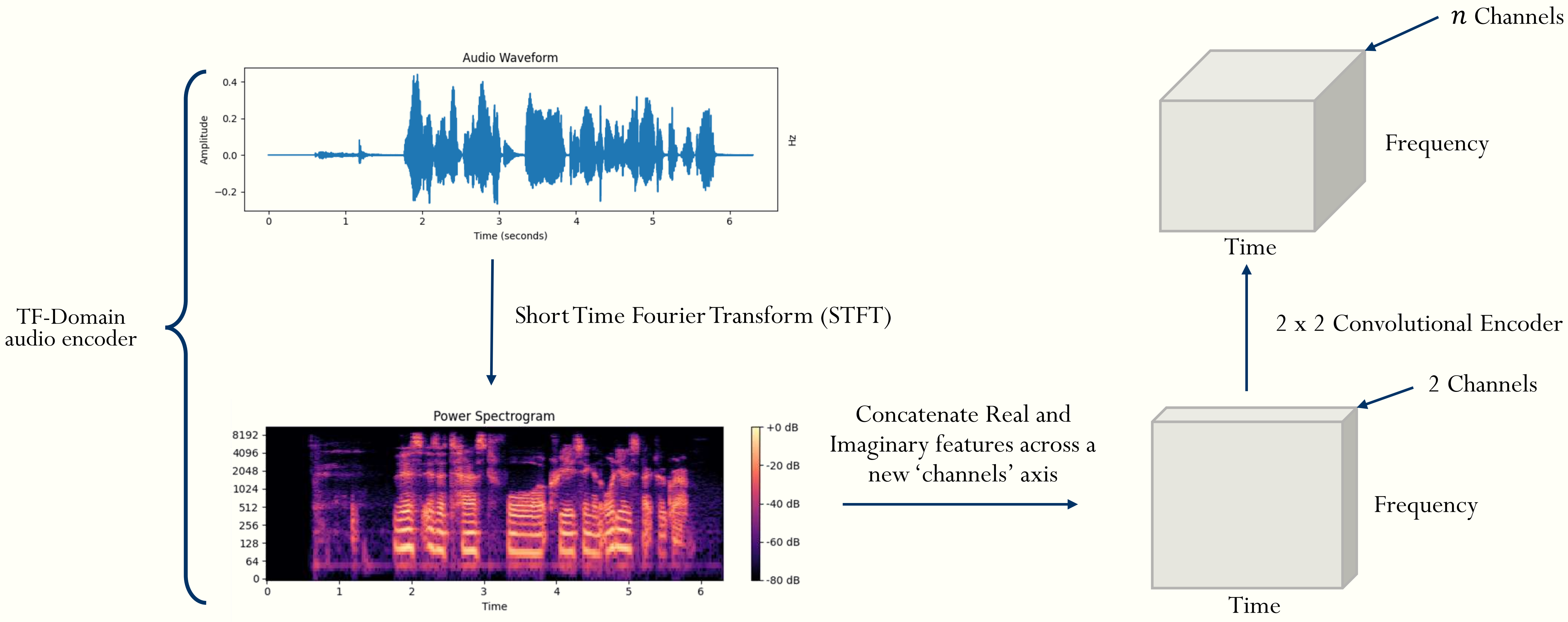
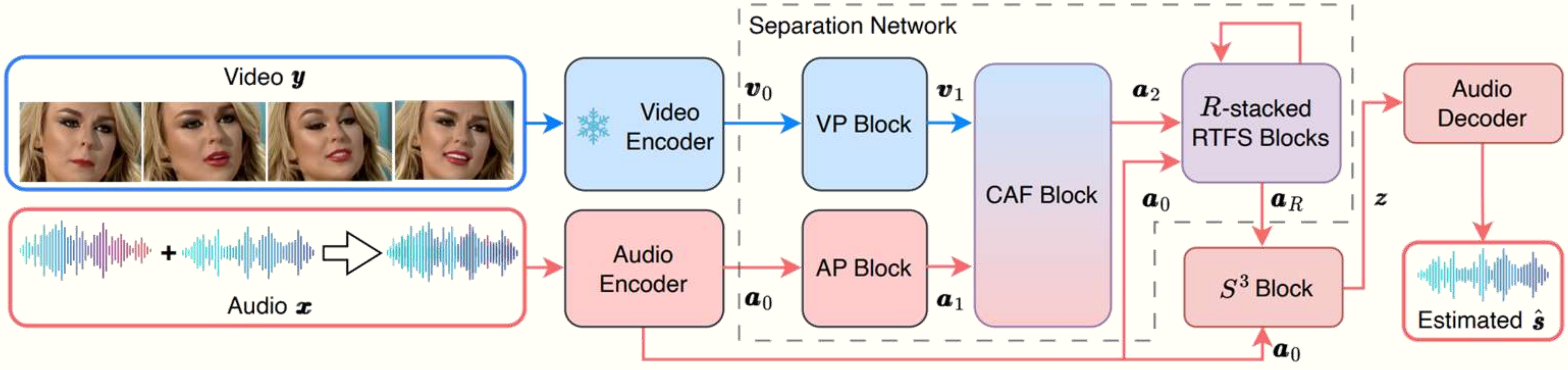


TABLE *of* CONTENTS

1. The Cocktail Party Problem
2. Motivations
3. RTFS-Net
4. Experimental Results
5. Conclusion

Methods: Target Speaker Extraction Pipeline



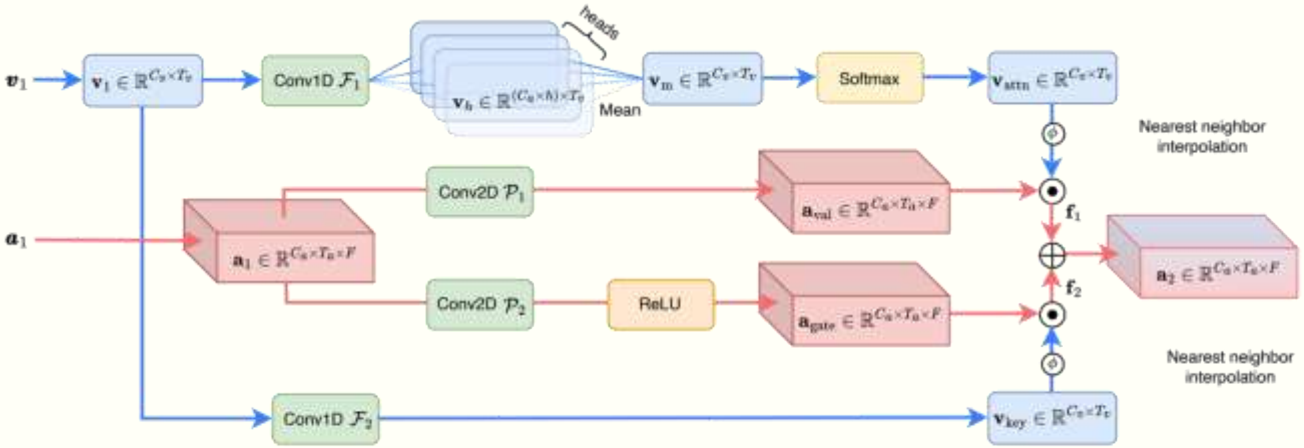
Key features:

- Pre-trained video encoder, Short-Time Fourier Transform (STFT) based audio encoder.
- Multimodal information fused using the **CAF Block**.
- R-Stacked **RTFS Blocks** distill the salient information.
- **S3 Block** extracts the target speaker's audio from the encoded audio mixture.
- Convolutional audio decoder utilizing the Inverse STFT.

Methods: Cross-Dimensional Attention Fusion Block Design

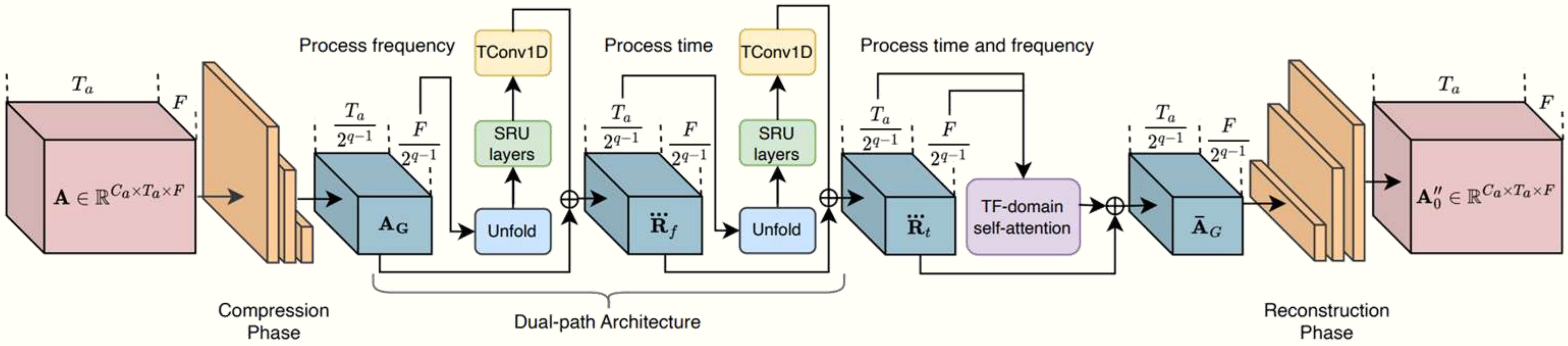
- Depth-wise, group convolution-based architecture.
- Designed to consume **as few resources as possible** while fusing the 2D visual data into the 3D audio data.
- Involves two fusion operations: the multi-headed **cross-attention** fusion f_1 and the **gated fusion** f_2.
- The attention fusion combines the information of multiple receptive fields via a **multi-headed attention operation** to filter the mixture of audio features using element-wise multiplication.
- The gated fusion again filters the mixture of audio features using the visual features as an **information gate.**

CAF Block



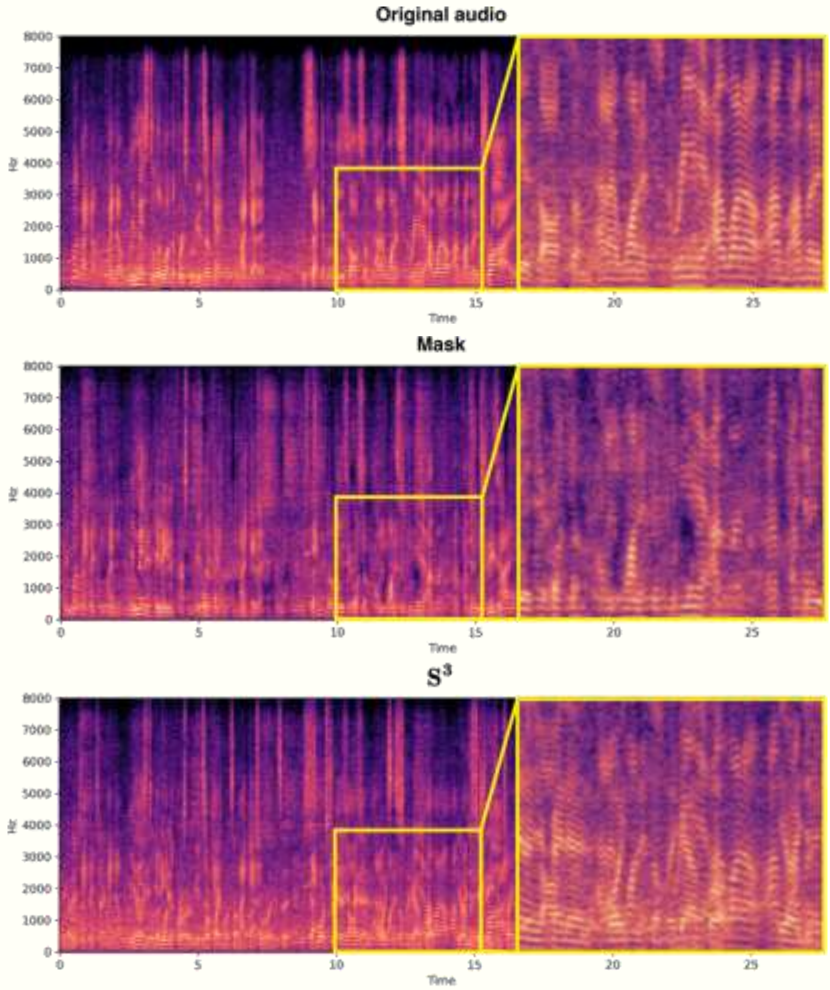
AV Fusion Strategies	LRS2-2Mix SI-SNRi	LRS2-2Mix SDRi	RTFS-Net Params (K)	RTFS-Net MACs (G)	Fusion Params (K)	Fusion MACs (M)
CTCNet Fusion (adapted)	11.3	11.7	528	14.3	197	6365
CAF Block (ours)	**11.7**	**12.1**	**339**	**8.0**	**7**	**83**

Methods: RTFS-Net Block Design



- Features are **compressed** to a more efficient size using **concentric convolutions** with stride 2.
- The frequency and time dimensions are processed **individually** using SRUs (optimized LSTMs), then in tandem using a **transformer** to capture inter-dependencies.
- Original dimensions are then **restored** using our Temporal-Frequency Attention Reconstruction (TF-AR) units.

Methods: Spectral Source Separation (S^3)



- Encoder produces **complex** features using the STFT.
- Real and imaginary features are concatenated, then passed to the convolutional encoder.
- Direct mask multiplication with the encoded mixture results in **critical information loss.**
- The S^3 Block solves this via a direct application of complex number properties.

Target Speaker Extraction Method	LRS2-2Mix SI-SNRi	LRS2-2Mix SDRi	RTFS-Net Params (K)	RTFS-Net MACs (G)	Extraction Params (K)	Extraction MACs (M)
Regression	10.0	9.9	**208**	**3.0**	**0**	**0**
Mask	10.8	11.2	224	3.6	16	534
Mask + DW-Gate	10.8	11.3	225	3.6	17	542
Mask + Gate	11.1	11.6	257	4.6	49	1595
S^3 (ours)	**11.3**	**11.7**	224	3.6	16	534

TABLE *of* CONTENTS

1. The Cocktail Party Problem
2. Motivations
3. RTFS-Net
4. **Experimental Results**
5. Conclusion

RESULTS

Model	LRS2-2Mix			LRS3-2Mix			VoxCeleb2-2Mix			Params	MACs	Time
	SI-SNRi	SDRi	PESQ	SI-SNRi	SDRi	PESQ	SI-SNRi	SDRi	PESQ	(M)	(G)	(ms)
CaffNet-C* 2021	-	12.5	1.15	-	12.3	-	-	-	-	-	-	-
Thanh-Dat 2021	-	11.6	-	-	-	-	-	-	-	-	-	-
AV-ConvTasnet 2019	12.5	12.8	-	11.2	11.7	-	9.2	9.8	-	16.5	-	60.3
VisualVoice 2021	11.5	11.8	2.78	9.9	10.3	-	9.3	10.2	-	77.8	-	130.2
AVLIT 2023	12.8	13.1	2.56	13.5	13.6	2.78	9.4	9.9	2.23	5.8	36.4	**53.4**
CTCNet 2022	14.3	14.6	**3.08**	17.4	17.5	3.24	11.9	13.1	**3.00**	7.0	167.2	122.7
RTFS-Net-4	14.1	14.3	3.02	15.5	15.6	3.08	11.5	12.4	2.94	**0.7**	**21.9**	57.8
RTFS-Net-6	14.6	14.8	3.03	16.9	17.1	3.12	11.8	12.8	2.97	**0.7**	30.5	64.7
RTFS-Net-12	**14.9**	**15.1**	3.07	**17.5**	**17.6**	**3.25**	**12.4**	**13.6**	**3.00**	**0.7**	56.4	109.9

- Comparison with existing methods on the LRS2-2Mix, LRS3-2Mix and VoxCeleb2-2Mix datasets.
- Larger SI-SNRi, SDRi and PESQ values indicate better performance.
- Lower parameter and MAC counts indicate smaller, more efficient models.
- RTFS-Net reduces computational complexity, and hence inference time significantly.

TABLE *of* CONTENTS

1. The Cocktail Party Problem
2. Motivations
3. RTFS-Net
4. Experimental Results
5. Conclusion

CONCLUSION

- RTFS-Net outperforms the prior SOTA method in both **inference speed** and **separation quality.**
- RTFS-Net **reduces** the number of parameters by **90%** and the number of MACs by **83%**.
- Efficiently fuse multimodal information using only 83 million MACs – a **97% reduction** (CAF Block).
- Effectively decode the separated audio without losing critical **amplitude** and **phase** information (S^3 Block).
- Apply powerful **RNN** and **Transformer** operations at compressed subspace to negate computational burden (RTFS Block).

Links

PAPER

CODE

DEMO

THANK
YOU

FOR YOUR
ATTENTION