MITSUBISHI ELECTRIC
Changes for the Better

Prompt-Conditioned Unified Audio Modeling: From Source Separation to Source-Aware Codecs

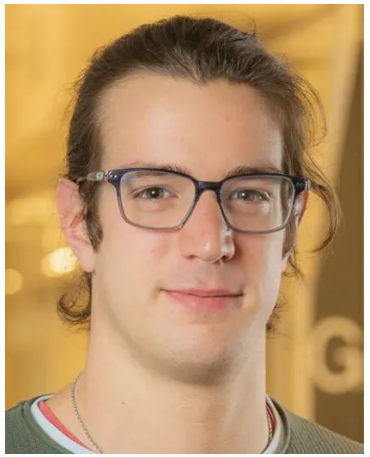
Jonathan Le Roux

March 26, 2026

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

Collaborators on today's topics

Kohei Saijo



Francesco Paissan

Yoshiki Masuyama

Ryo Aihara

Gordon Wichern

Francois Germain

Janek Ebbers

Zexu Pan

Jiangyu Han
(BUT)

Marc Delcroix
(NTT)

The Cocktail Party Problem

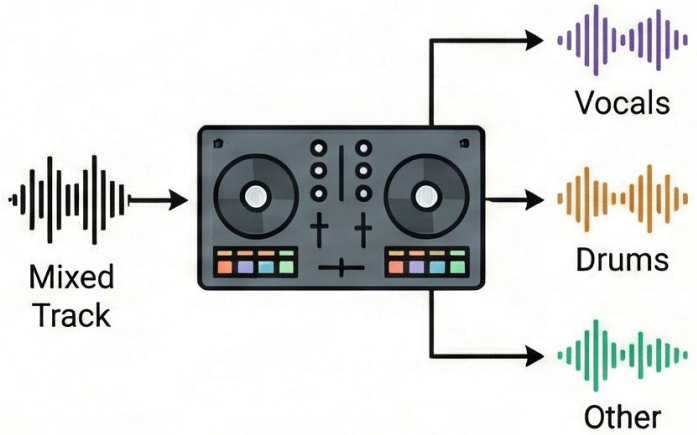
Generated by ChatGPT

Many applications

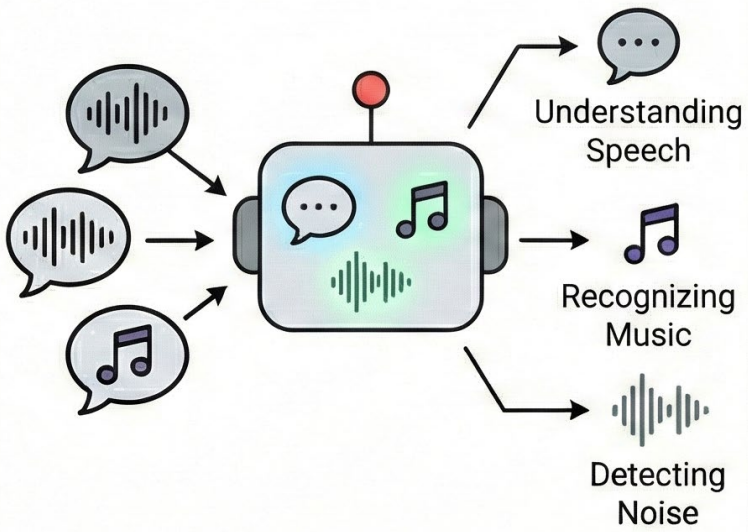
Images courtesy of Kohei Saijo and ChatGPT

Hearing Aid

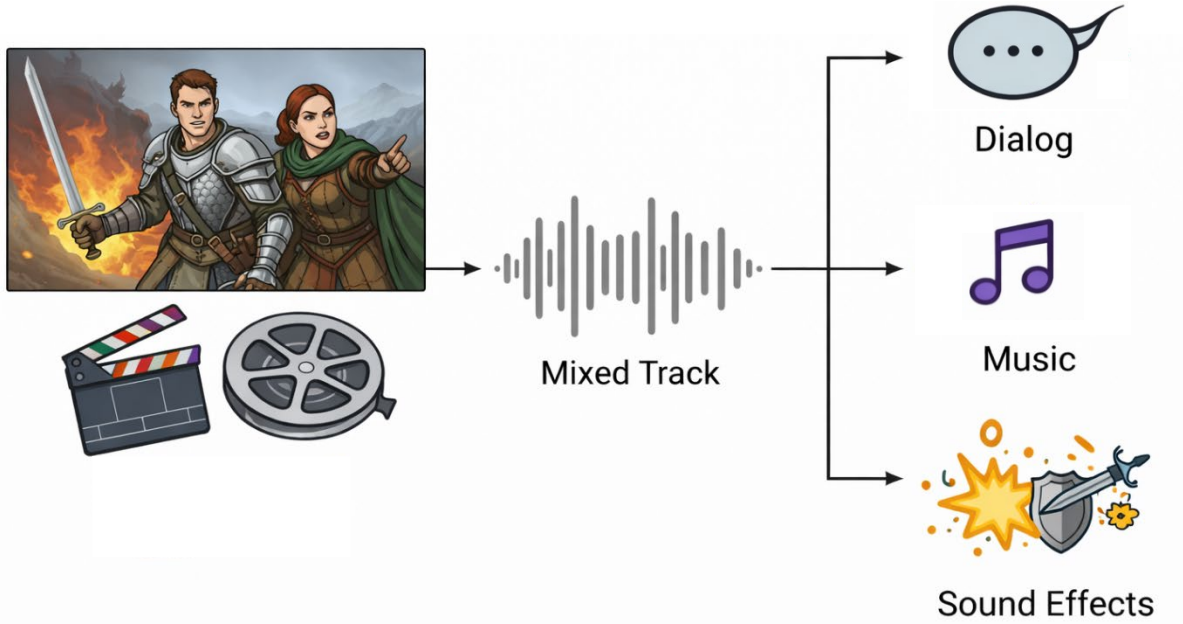
Meeting Separation



Sound Remixing



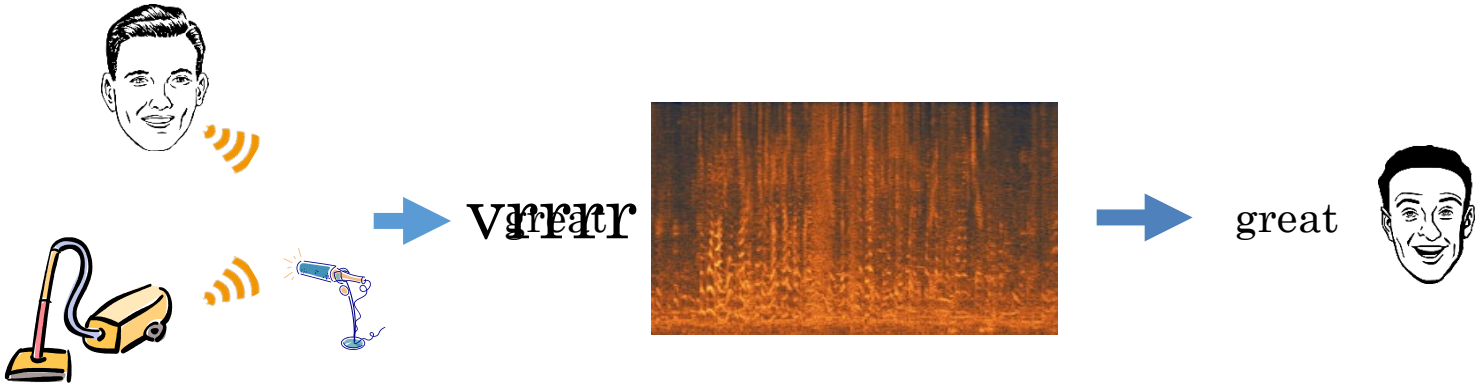
Robot Audition



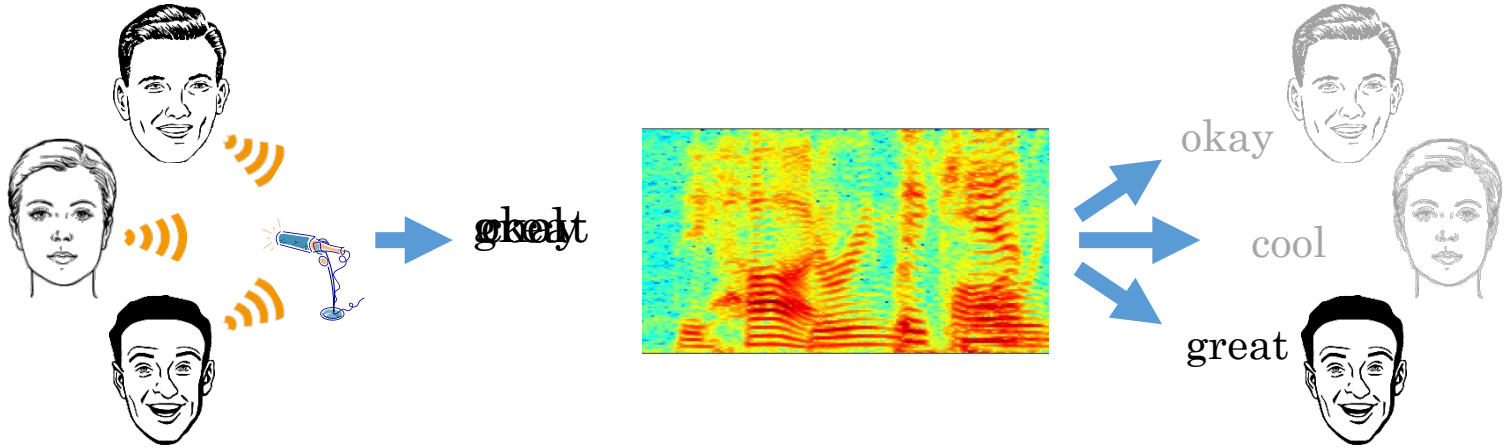
Cinematic Sound Separation

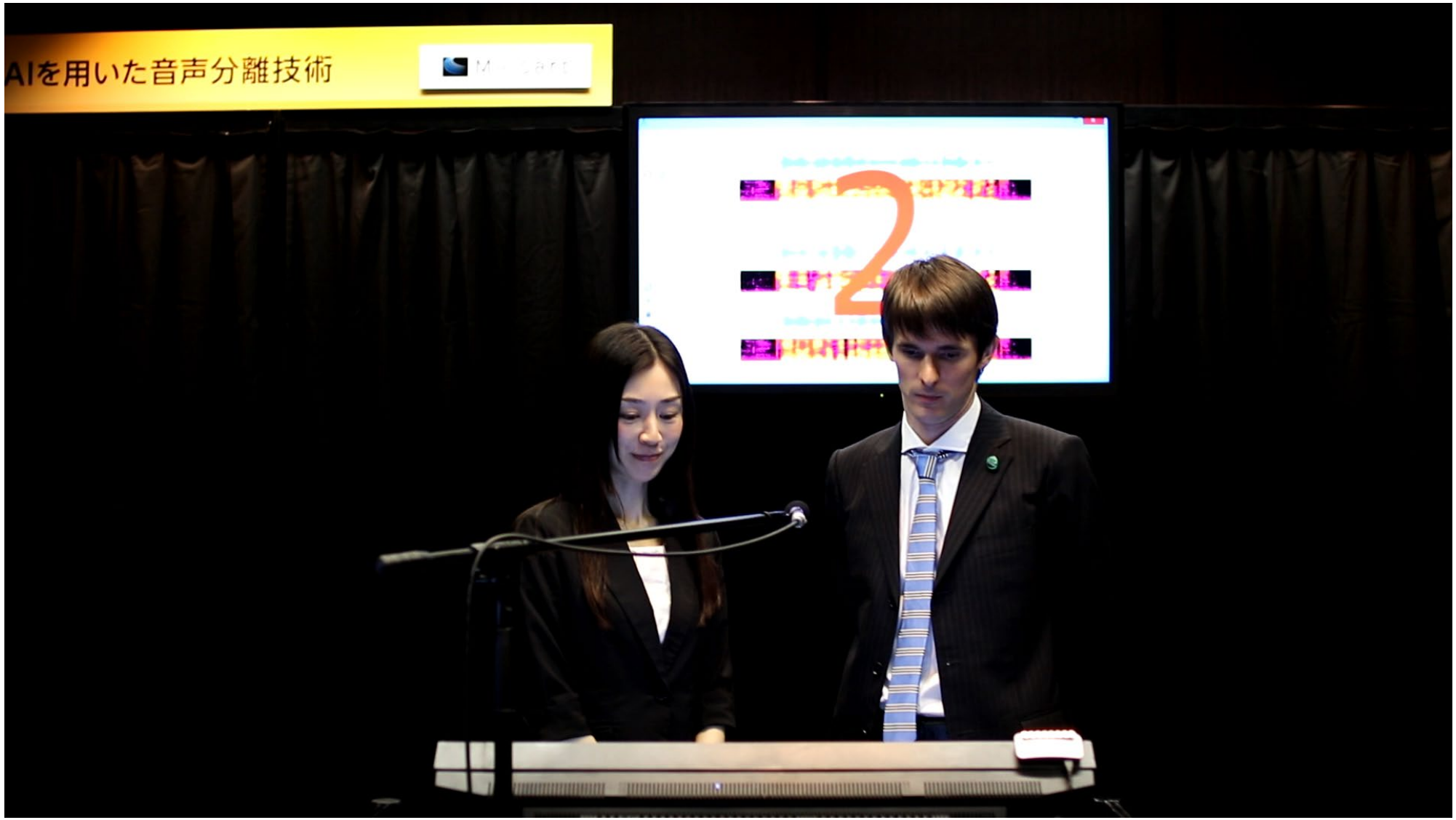
Speech Enhancement and Separation

Enhancing speech in challenging noise environments using a single microphone



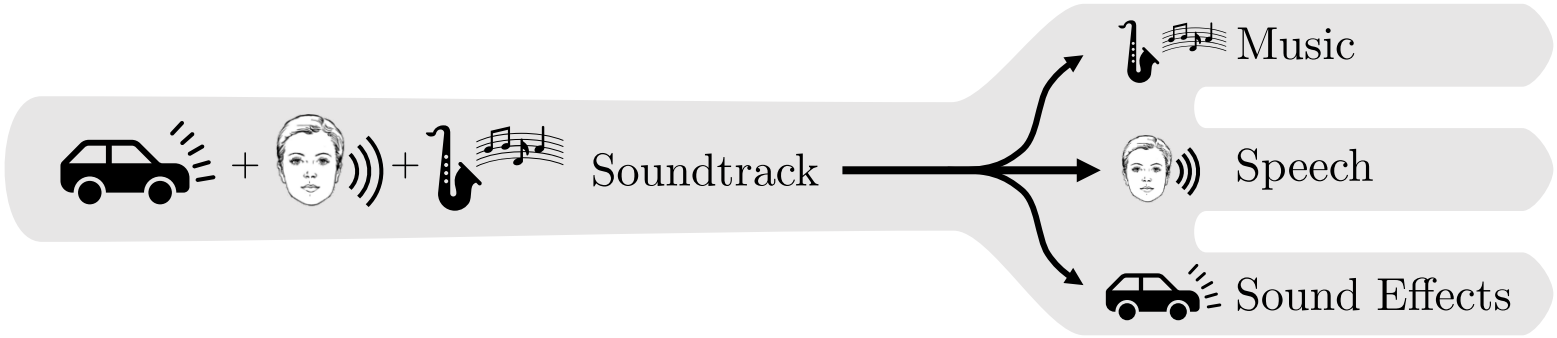
Separation of multiple unknown speakers using a single microphone

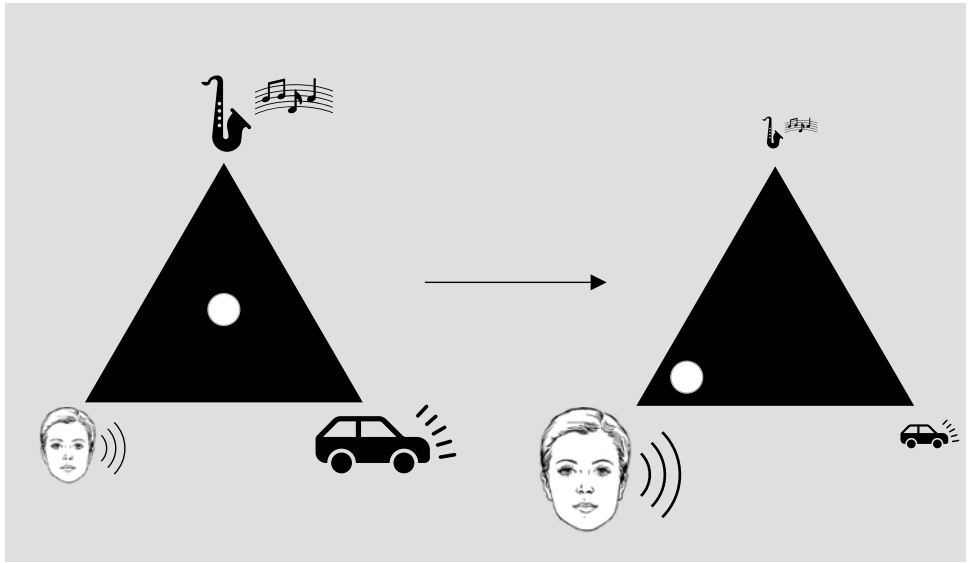


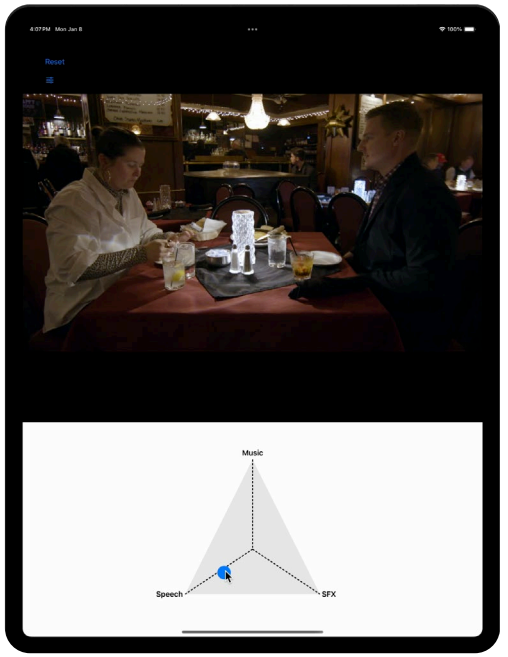


From the Cocktail Party Problem to the Cocktail Fork Problem

Three-Stem Audio Separation For Real-World Soundtracks

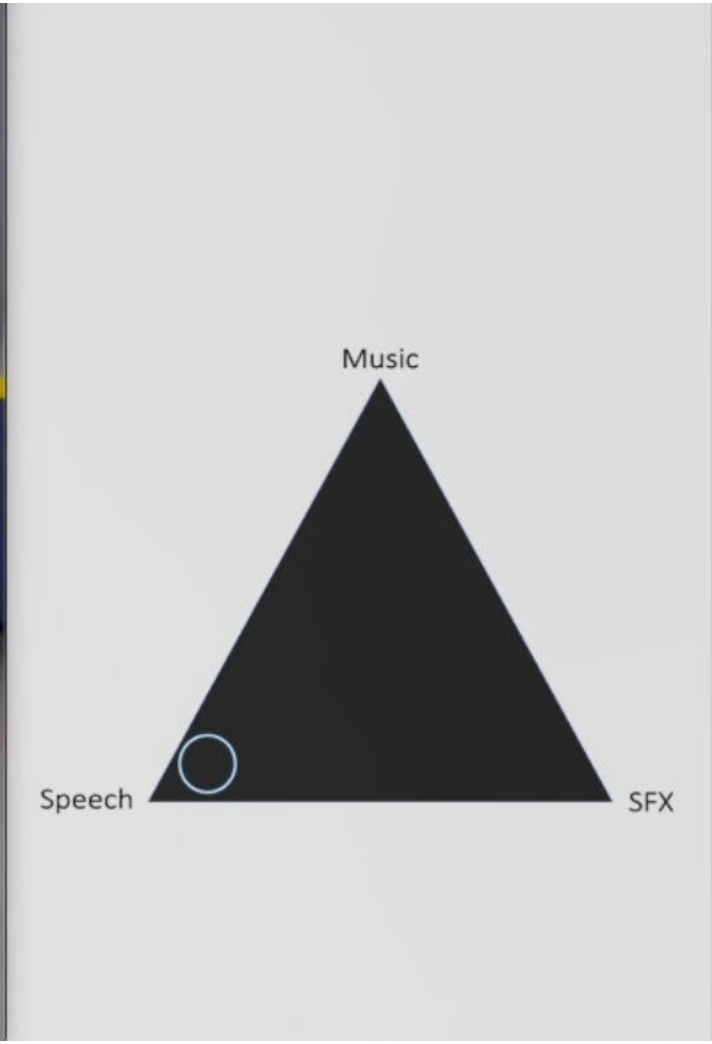






Showcased at the Mitsubishi Electric Booth @CES 2024

MITSUBISHI ELECTRIC
Changes for the Better



Can we handle all separation tasks using a single model?

- Why do we care? Now we have SAM Audio!
- SAM Audio is great, but it solves a different problem: Target Source Extraction
- Target source extraction can’t solve everything
  - Not trivial to indicate context
  - Not easy to separate multiple sources of a similar type, i.e., no **instance-level** separation
    - E.g., two male speakers
  - One source (or group of sources) at a time

Can we handle all separation tasks using a single model?

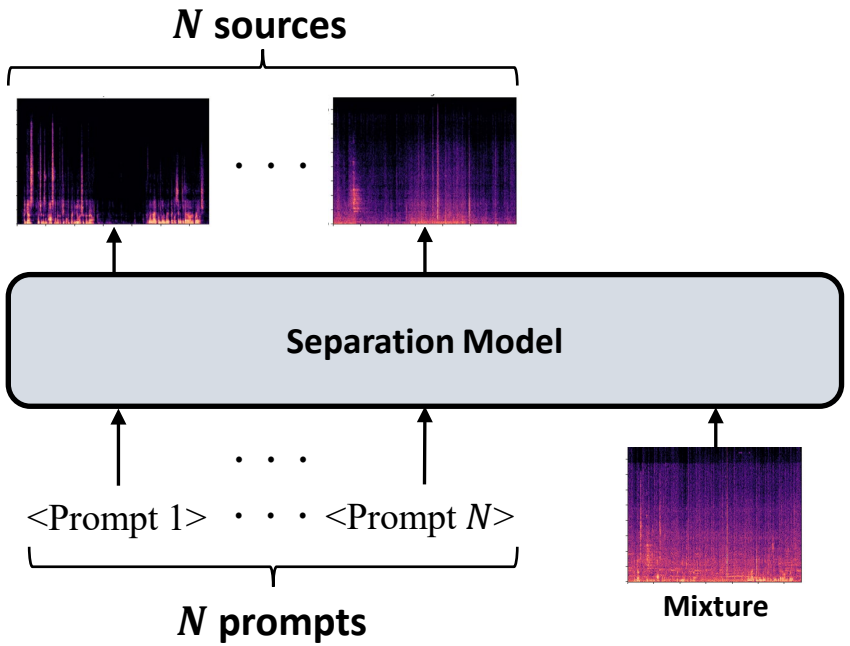
- **Different applications (or tasks) have different target sources of interest**
  - Task-specific models need to be deployed for each application, which is not efficient

Task	Sources of interest
Speech enhancement (SE)	Speech, Noise
Speech separation (SS)	Speech $\times$ N, Noise
Environmental sound separation (USS)	Sound effects (SFX) $\times$ N
Music source separation (MSS)	Vocals, Bass, Drums, Other inst.
Cinematic audio source separation (CASS)	Speech, SFX-mix, Music-mix

- **NN's powerful modeling capability may enable unification of all separation tasks**
  - LLMs handle various tasks that were originally handled by specialist models
  - To address all separation tasks, the model needs to handle
    1) arbitrary classes of sources,
    2) a variable number of sources,
    3) with an explicit control of granularity

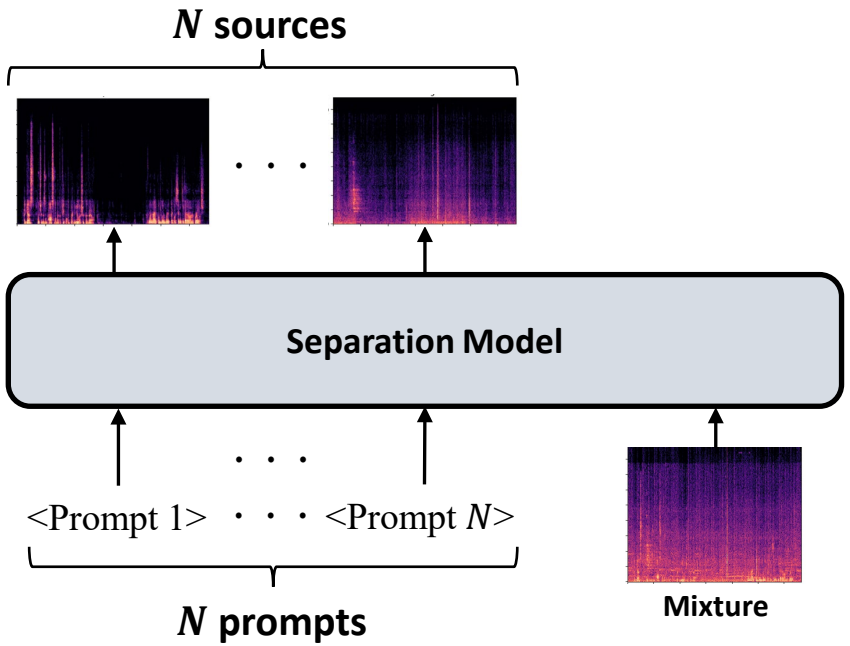
How Can We Build a Unified Source Separation Model?

- **Requirements**
  1. A conditional model which can change its behavior in inference
  2. A model that accepts a variable number of prompts
  3. A model that accepts multiple identical prompts



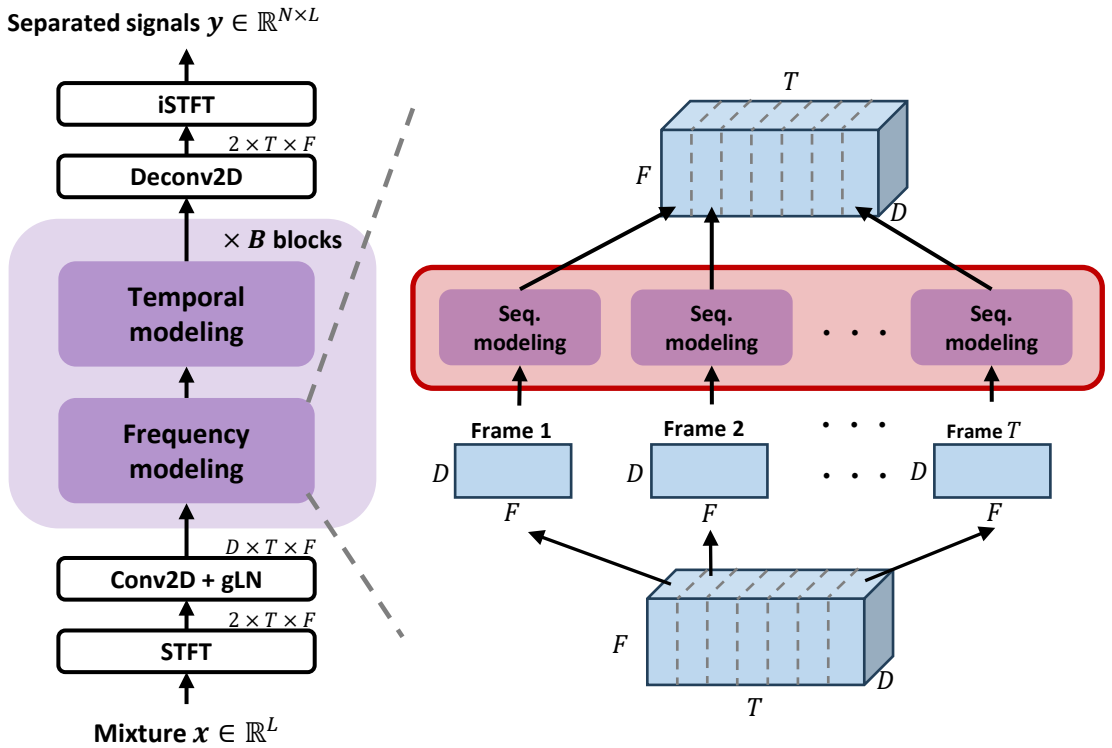
How Can We Satisfy the Requirements?

- **Transformer-based separation models are good candidates**
  1. A conditional model which can change its behavior in inference
     → The model can change its behavior by prompting
  2. A model that accepts a variable number of prompts
     → Transformers work regardless of the input sequence length
  3. A model that accepts multiple identical prompts
     → Positional encodings make prompts different from each other



TF-domain Dual-Path Separation Models

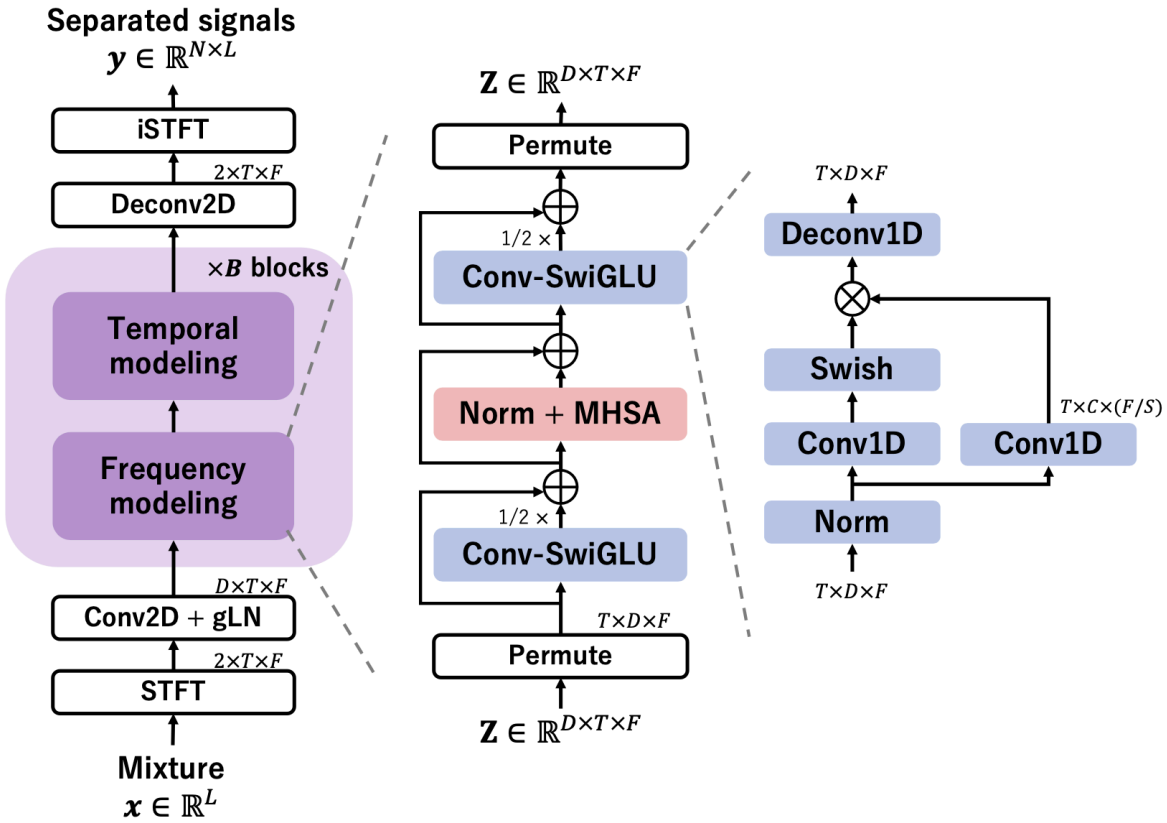
- **Base of the current state-of-the-art separation model**
  - Alternate sequence modeling along time and frequency dimensions
  - LSTM is very strong [Wang+, 2023], but it would be great to have a Transformer alternative
    - Scalability, prompting, etc.



[Wang+, 2023]: Z.-Q. Wang et al., "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM TASLP*, 2023.

TF-Locoformer [Saijo+, IWAENC2024 Best Paper Award Candidate]

- **TF domain Transformer with LOcal modeling by COnvolution**
  - A design inspired by Conformer [Gulati+, 2020] and Transformer++ [Gu+, 2024] † models

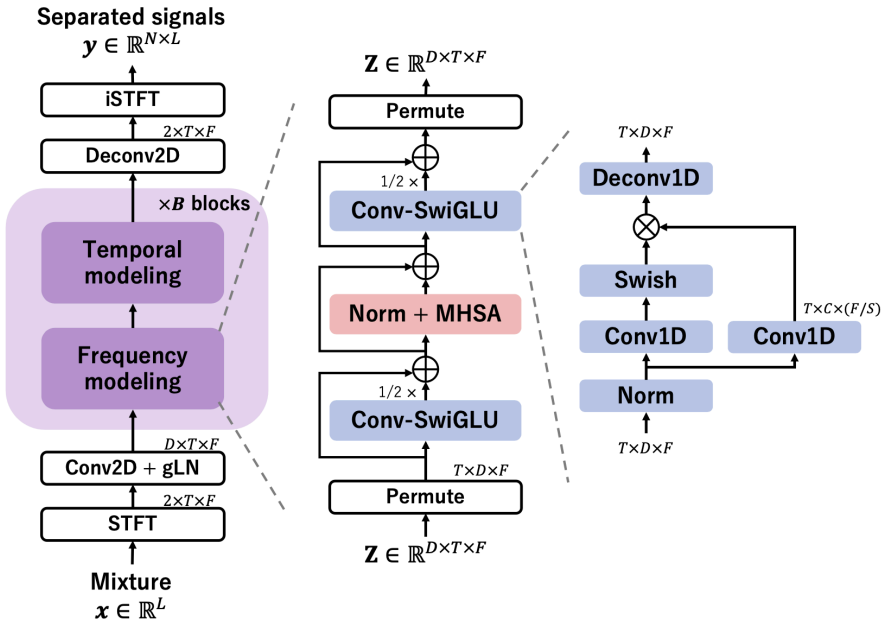


[Gulati+, 2020]: A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," *Interspeech,* 2020.

†: Existing model named in [Gu+, 2024]: A. Gu et al., "Mamba: Linear-time sequence modeling with selective state spaces." *First conference on language modeling*. 2024.

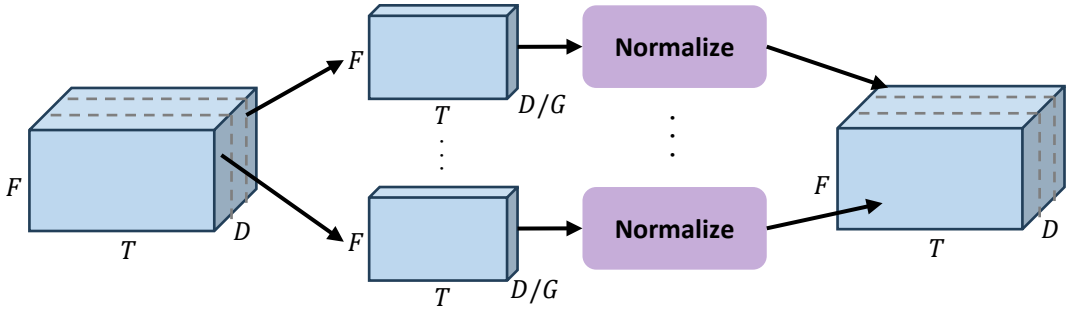
Key Components of TF-Locoformer

- **Conv-SwiGLU FFN**
  - Convolutional layers inspired by Conformer
  - SwiGLU activation inspired by Transformer++



- **Macaron-style architecture**
  - Two FFNs before and after MHSA

- **RMSGroupNorm**
  - Split D-dimensional vector into G groups and normalize each D/G-dimensional vector
  - This may encourage disentanglement of each source's features



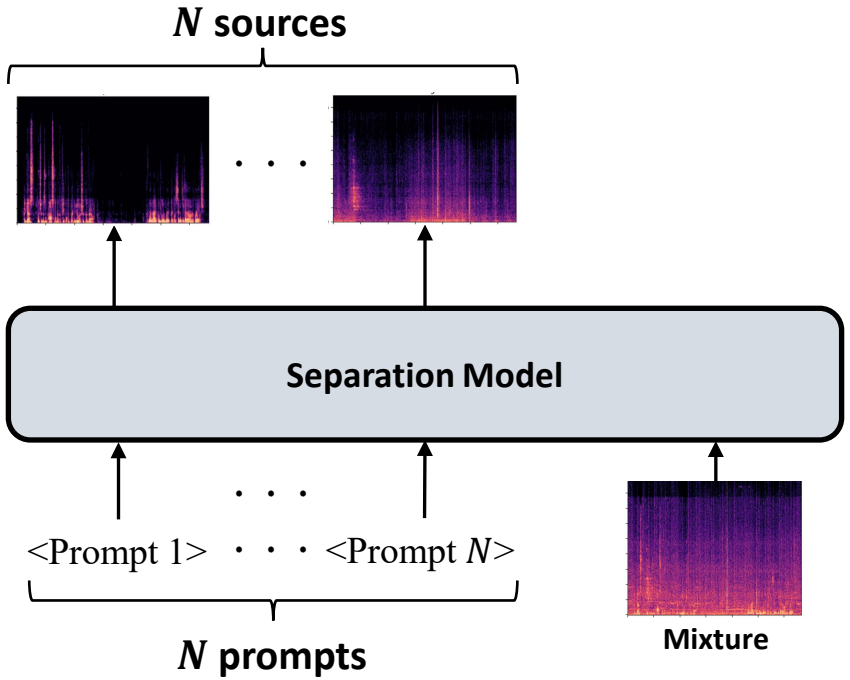
Speech Separation Experiments

- **Dataset:** WSJ0-2mix (anechoic) and WHAMR! (noisy reverberant)
- **Metric:** SI-SDR [dB]
- **Results:**
  - Comparable performance to LSTM-based model, TF-GridNet
  - Better performance with larger model

Model	#params	WSJ0-2mix	WHAMR!
TF-GridNet (S)	5.5 M	-	17.1
TF-GridNet (M)	14.4 M	23.5	-
TF-Locoformer (S)	5.0 M	22.0	17.4
TF-Locoformer (M)	15.0 M	23.6	18.5
TF-Locoformer (L)	22.5 M	24.2	-

Unified Source Separation based on TF-Locoformer

- TF-Locoformer satisfies the requirements to build prompting-based models
  - Sequence-length invariant
  - But can make prompts different from each other thanks to the positional encoding



Prompts to be Considered

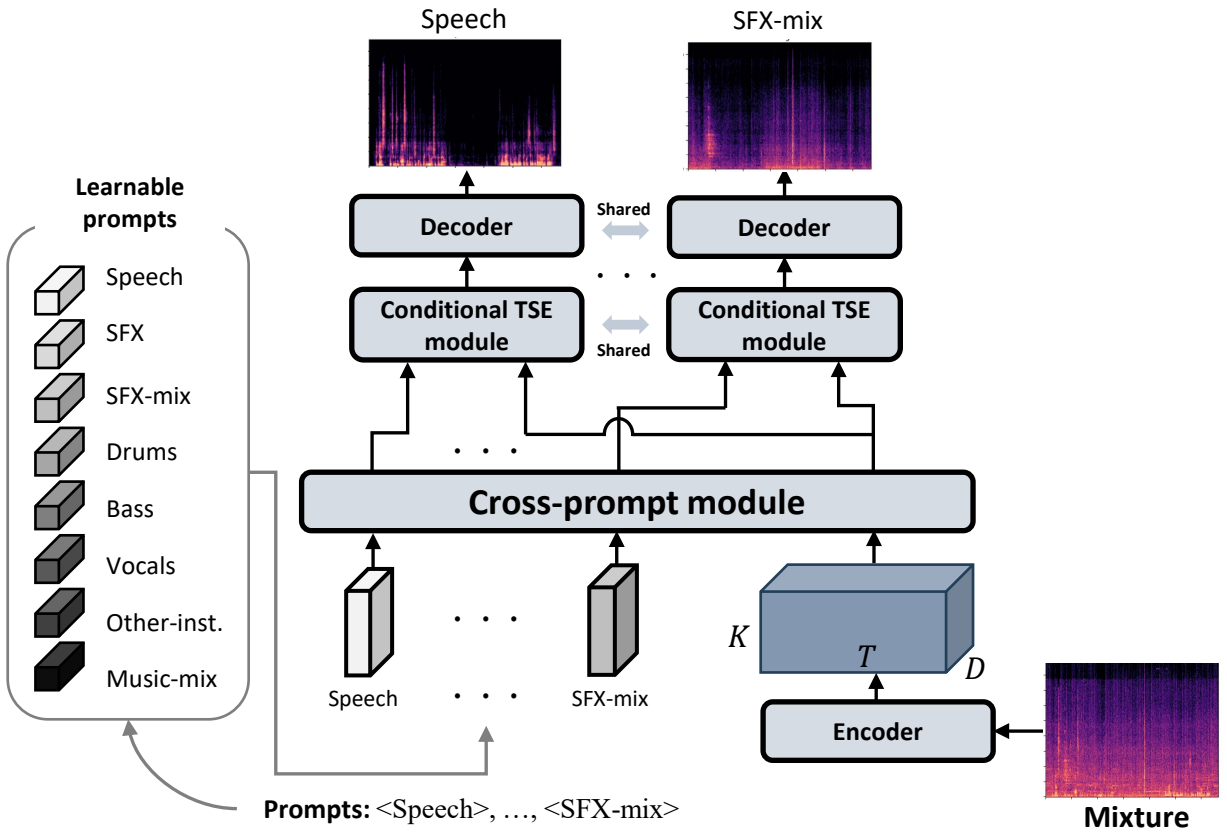
- **8 types of prompts**
  - Speech: <Speech>
  - Sound effects: <SFX-mix> <SFX>
  - Music: <Music-mix> <Drums> <Bass> <Vocals> <Other inst.>

- **Major tasks can be covered by changing the combination of prompts**

Task	Prompts
Speech enhancement (SE)	<Speech>, <SFX-mix>
Speech separation (SS)	<Speech> x *N*, <SFX-mix>
Environmental sound separation (USS)	<SFX> x *N*
Music source separation (MSS)	<Drums>, <Bass>, <Vocals>, <Other inst.>
Cinematic audio source separation (CASS)	<Speech>, <SFX-mix>, <Music-mix>

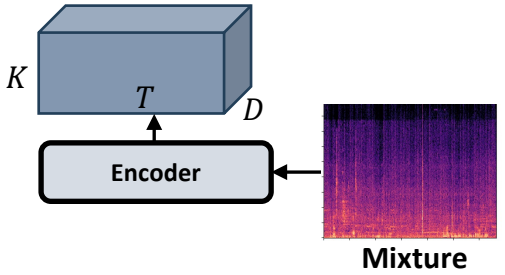
Task-aware Unified Source Separation (TUSS) [Saijo+, ICASSP2025]

- **Model satisfies the requirements by using self-attention**



1. Encoder

- **STFT-domain band-split module** **[Luo+, IEEE/ACM TASLP2023]**
  - Applies STFT to the mixture waveform $\boldsymbol{X} \in \mathbb{R}^{2\times F\times T}$
  - Further encodes the spectrogram into $\boldsymbol{Z} \in \mathbb{R}^{D\times K\times T}$

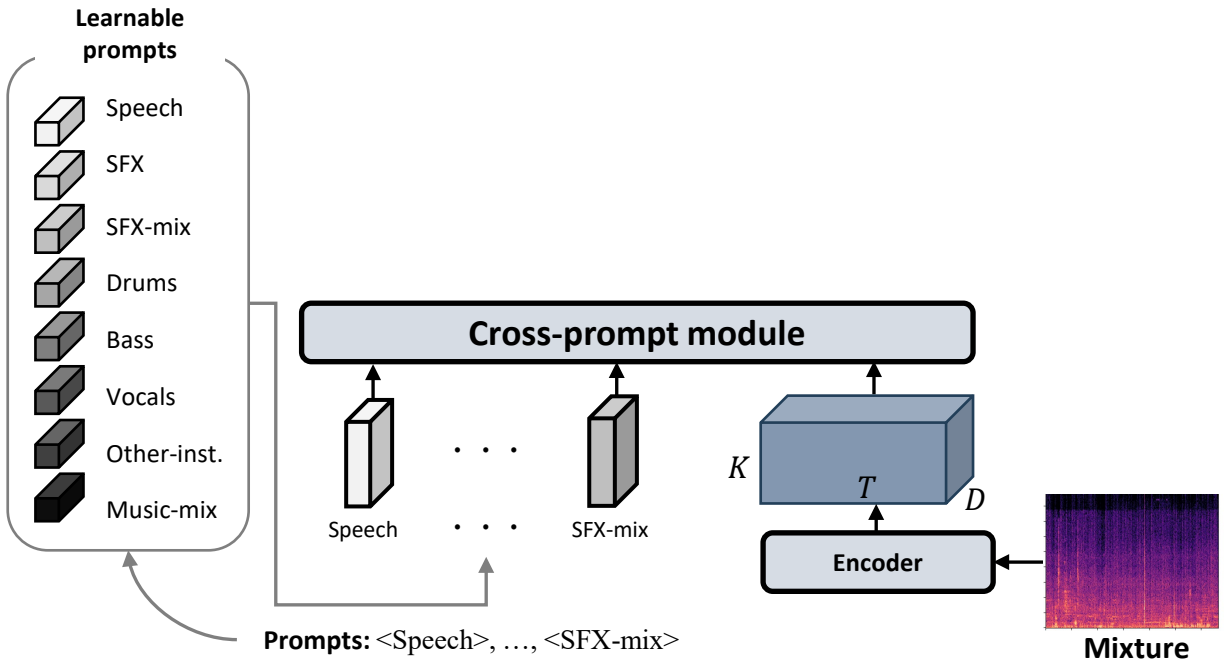


[Luo+, 2023]: Y. Luo and J. Yu, "Music source separation with band-split RNN," *IEEE/ACM TASLP*, 2023.

2. Cross-prompt Module

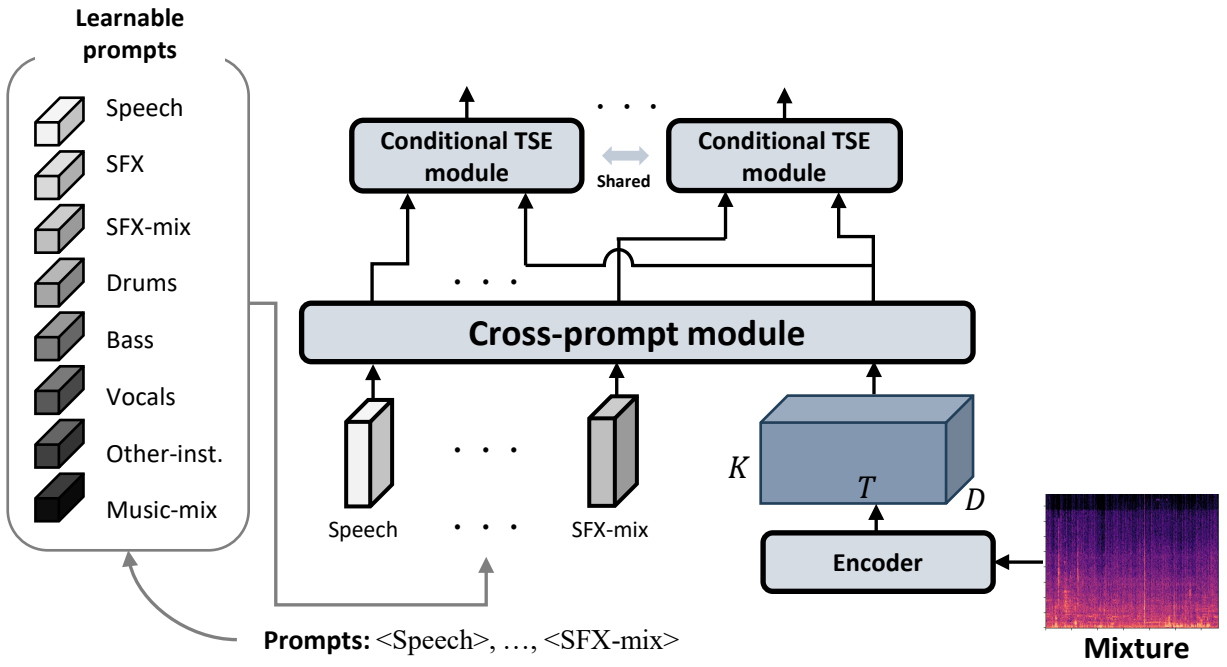
Mixes the prompts and the encoded features via self-attention

- The prompts and the encoded features are conditioned on each other
- Enables us to use a variable number of prompts and multiple identical prompts



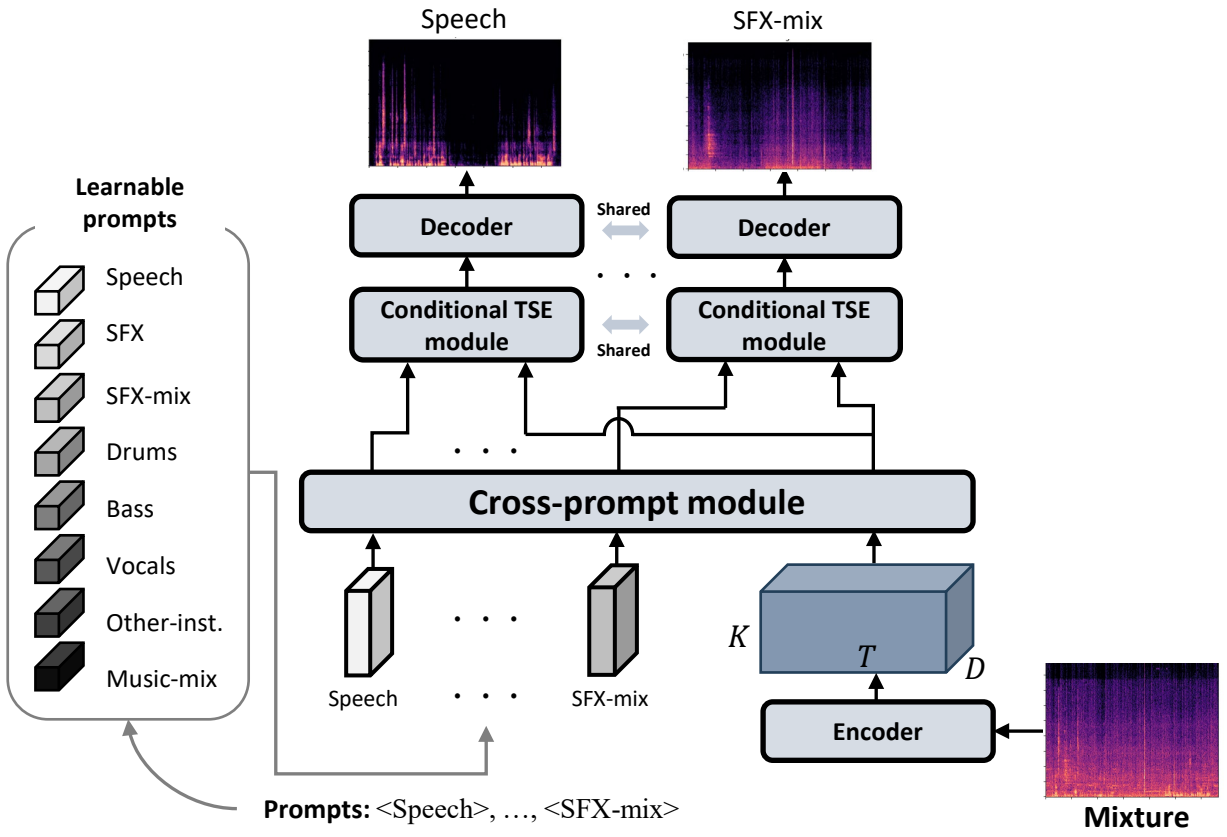
3. Conditional TSE Module

- **Processes each pair of a prompt and the features one by one**
  - Conditioning by element-wise product: $\boldsymbol{Z_n} = \boldsymbol{P_n} \odot \boldsymbol{Z} \in \mathbb{R}^{D \times K \times T}$
  - Further applies some TF-Locoformer blocks
  - Variable number of prompts is acceptable as the TSE module is shared for all n



4. Decoder

- **Band-wise decoding to project the separated feature back to the STFT domain**
  - Inverse transformation of the encoder
  - Applied independently to each source



Experimental Setup

- **Training Data: on-the-fly mixing of audios from collection of public data**
  - Randomly sample **2-4 prompts** and the corresponding audio data, and mix them

Category	Datasets
Speech	VCTK, WSJ, LibriVox
SFX	FSD50K
SFX-mix	WHAM! DEMAND, FSD50K
Music Inst.	MUSDB-HQ, MOISESDB
Music-mix	FMA, MUSDB-HQ, MOISESDB

- **Validation/testing data: public benchmarks for each task**
  - Voicebank-DEMAND (SE), WHAM! (SS), FUSS (USS), MUSDB-HQ (MSS), DnR (CASS)

Unified Source Separation Experiments

- **Methods:**
  - Unconditional: unconditional separation model with fixed number of outputs
  - TUSS: proposed TUSS-based conditional model
- **Metrics:**
  - SI-SDR [dB], except for MSS where we use SNR [dB]
- **Results:**
  - Prompting-based unified model outperforms the unconditional model
  - Successfully handled tasks that require different #sources and granularity

Method	SE	SS	USS	MSS	CASS	Average
Unconditional	14.0	6.8	8.1	4.9	6.0	8.0
TUSS	**14.8**	**9.1**	**9.6**	**6.8**	**9.1**	**9.9**

Comparison with Specialist Models

- **Two types of models trained on different data**
  - Specialists: multiple models, each trained on all the data **for a given task**
  - TUSS: one model trained on all the data available **for all tasks**
- **Results:**
  - Medium unified model could not outperform the specialist models
  - Generally, larger model benefits from larger data, what if we scale the model up?

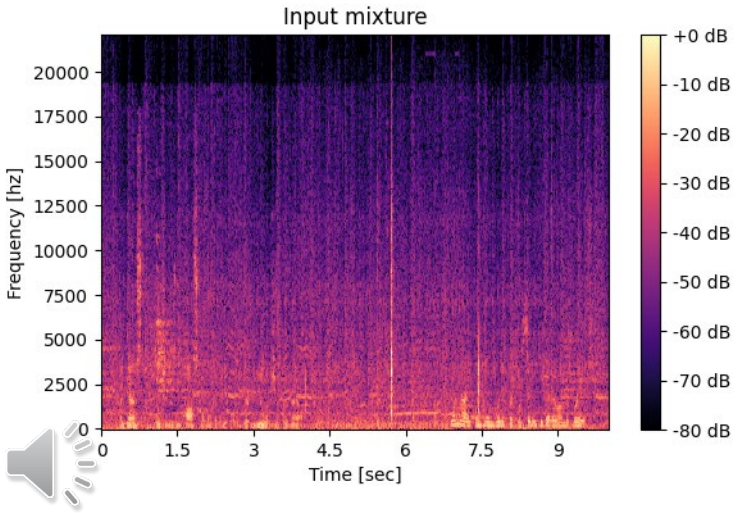
Type	SE	SS	USS	MSS	CASS	Average
Specialist	**15.9**	**10.3**	**10.2**	**8.3**	**9.7**	**10.9**
TUSS	14.8	9.1	9.6	6.8	9.1	9.9

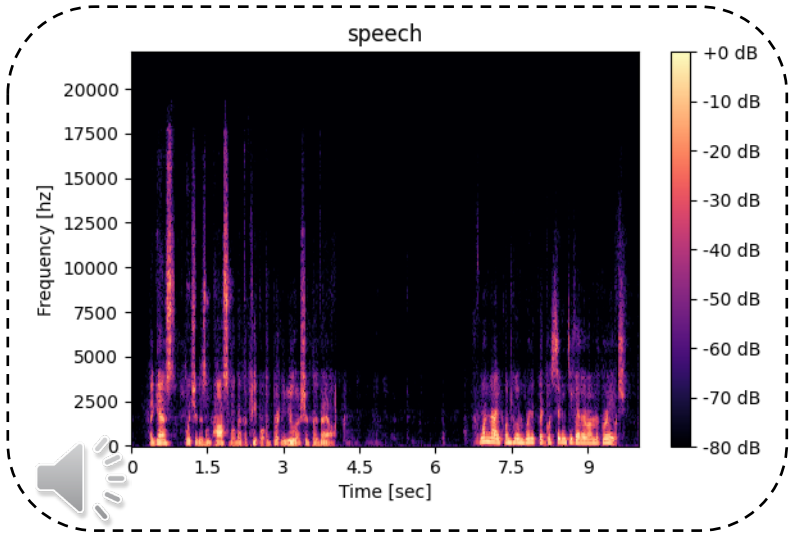
Results on Larger Model

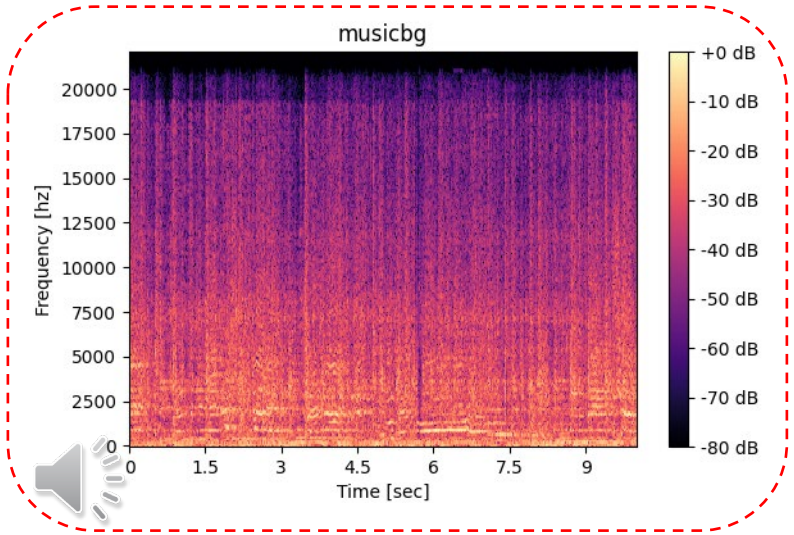
- **TUSS achieved comparable performance to the specialists on several tasks**
  - As expected, larger model benefits from large data
  - Showed its potential to serve as a foundation model for source separation

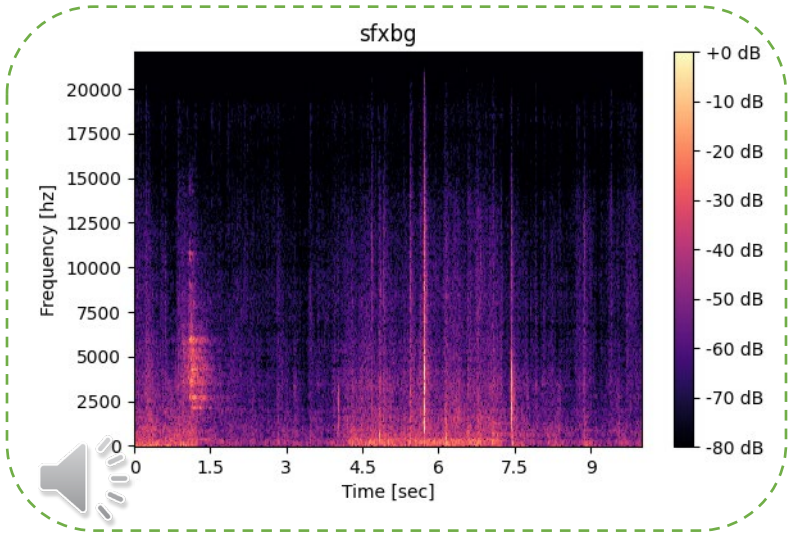
Model size	Type	SE	SS	USS	MSS	CASS	Average
Medium	Specialist	15.9	10.3	10.2	8.3	9.7	10.9
	TUSS	14.8	9.1	9.6	6.8	9.1	9.9
Large	Specialist	**16.0**	**11.4**	10.0	**9.1**	10.0	11.3
	TUSS	15.1	10.3	**12.2**	7.4	**10.1**	11.0

TUSS output with prompts: <Speech>, <Music-mix>, <SFX-mix>

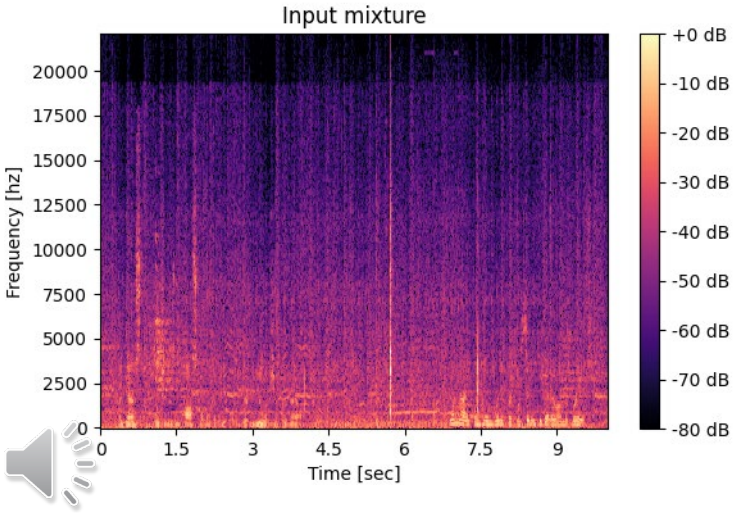


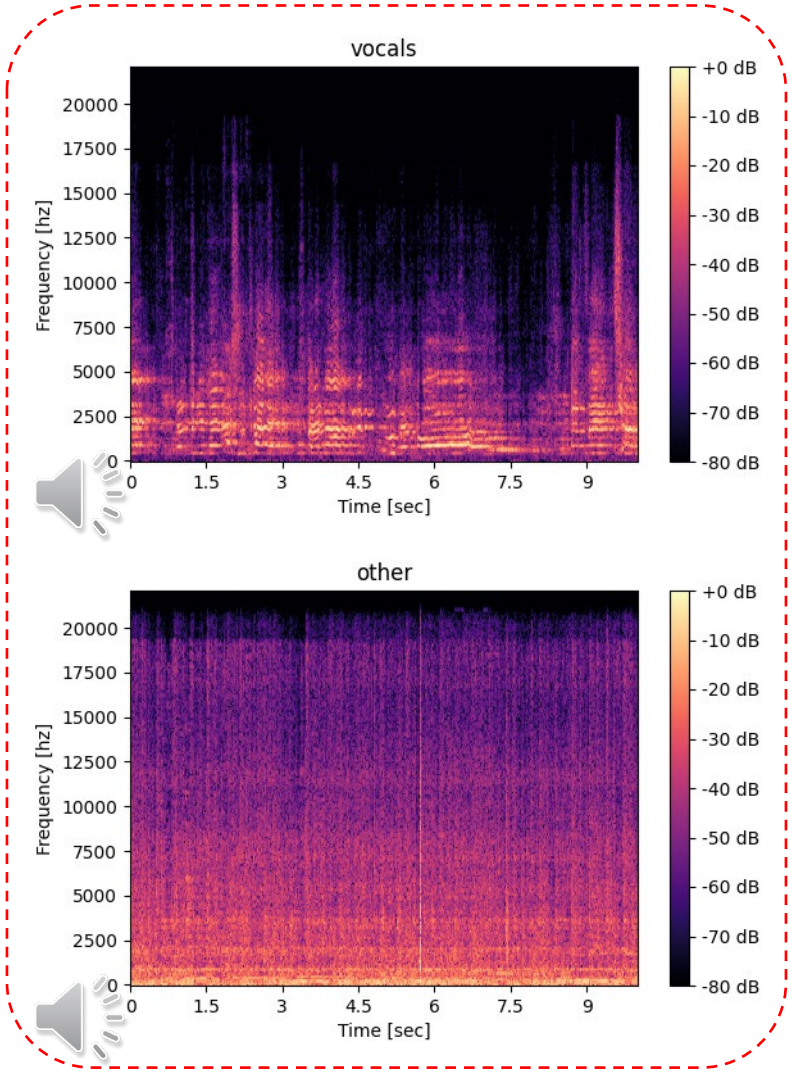


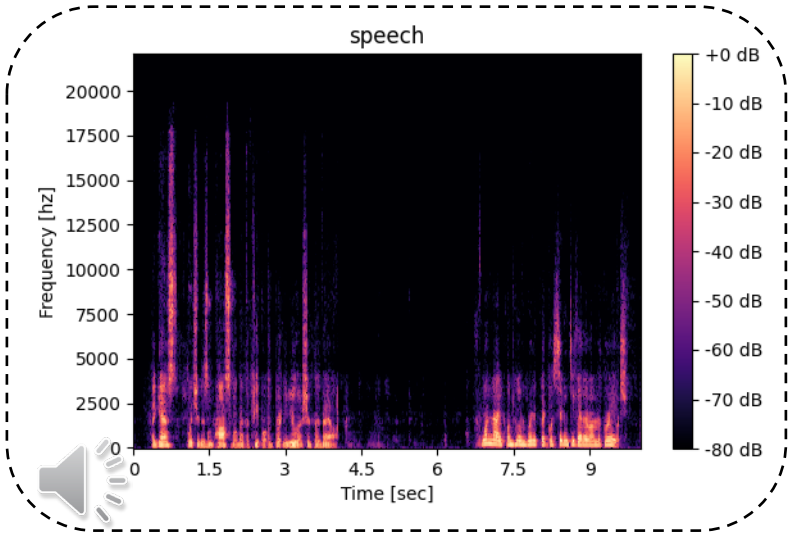


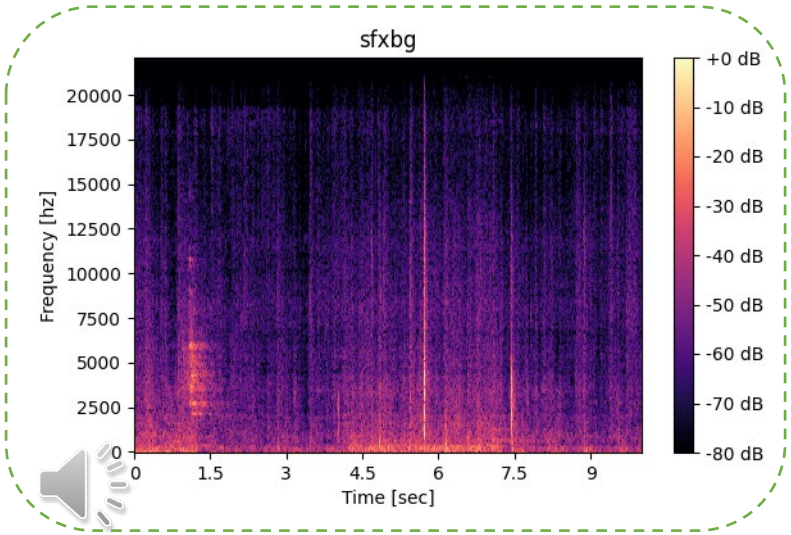


TUSS output with prompts: <Speech>, <Vocals>, <Other inst.>, <SFX-mix>

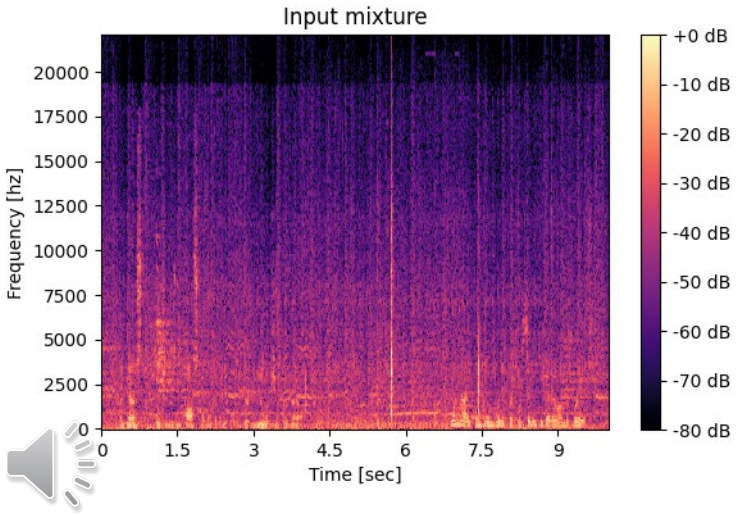


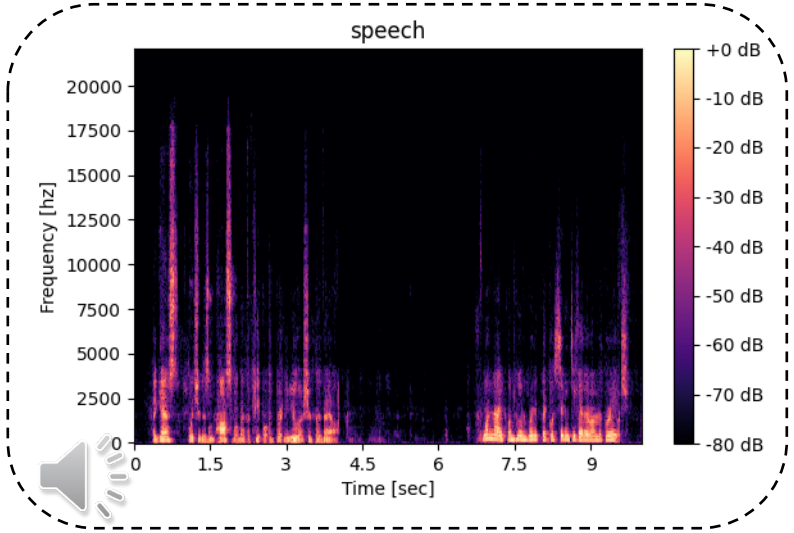


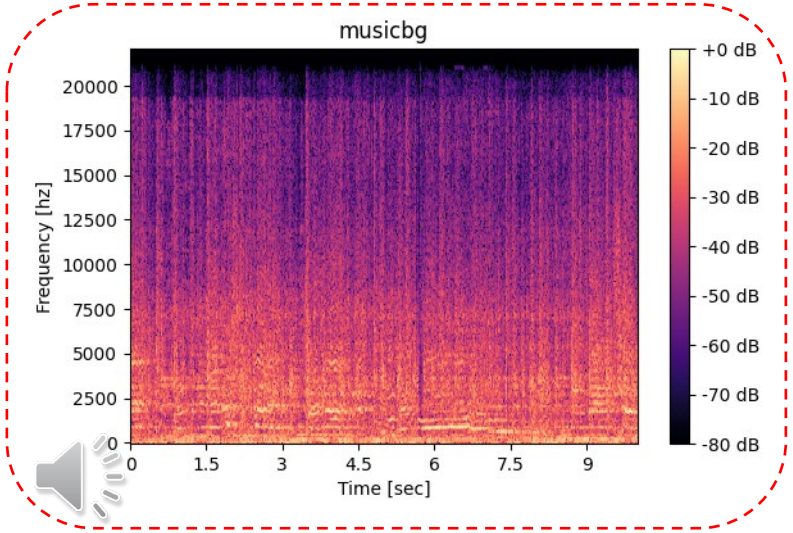


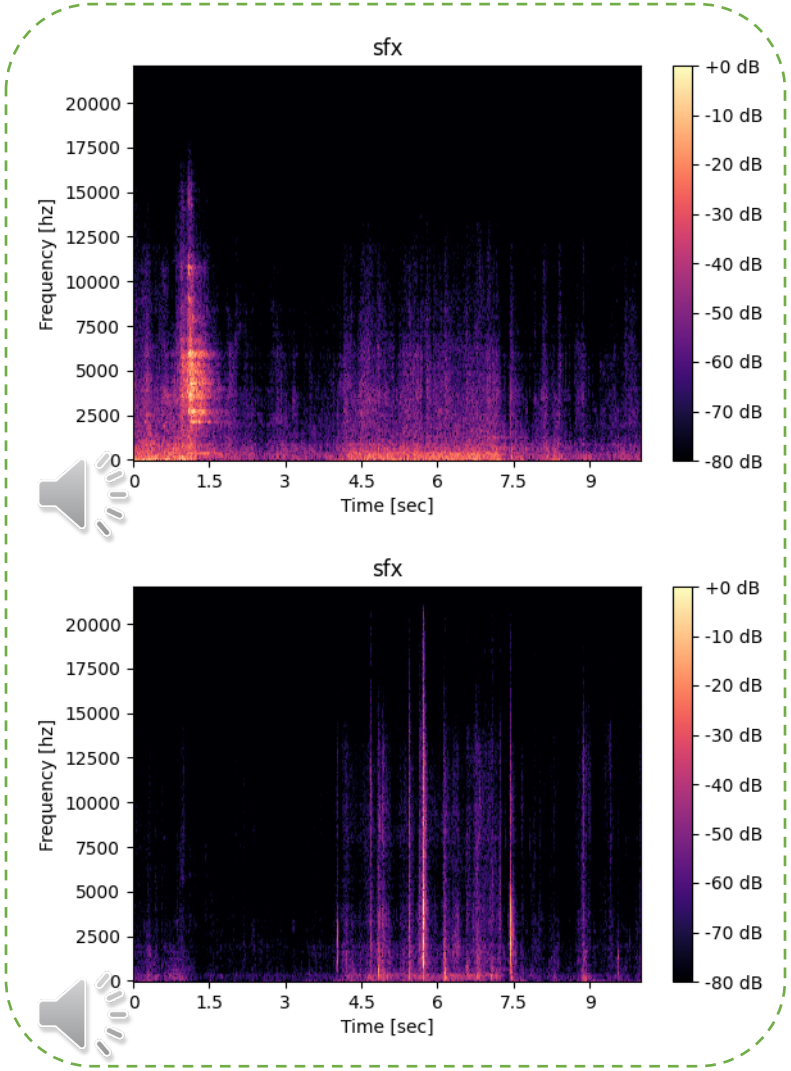


TUSS output with prompts: <Speech>, <Music-mix>, <SFX>, <SFX>

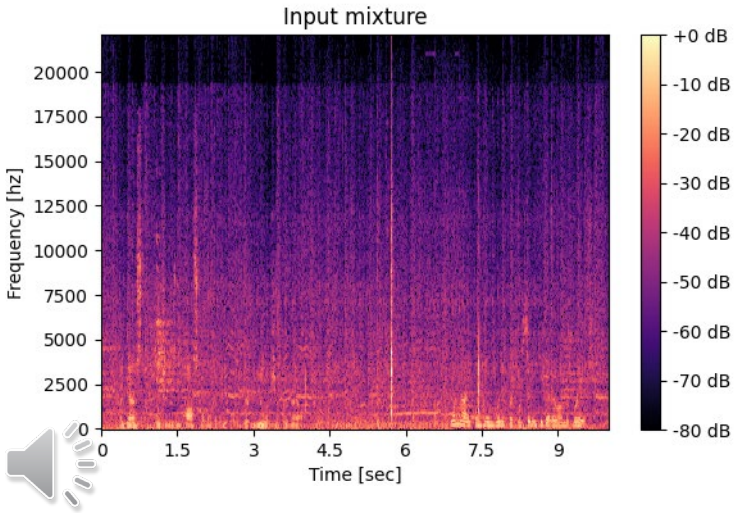


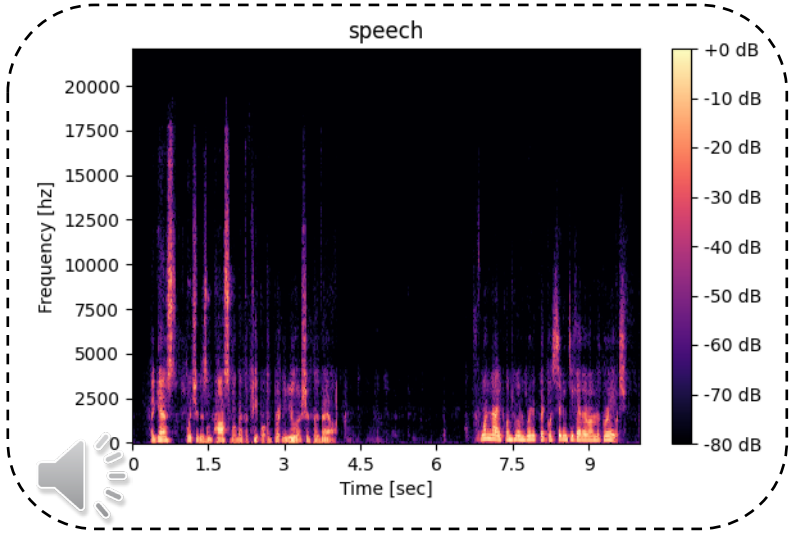


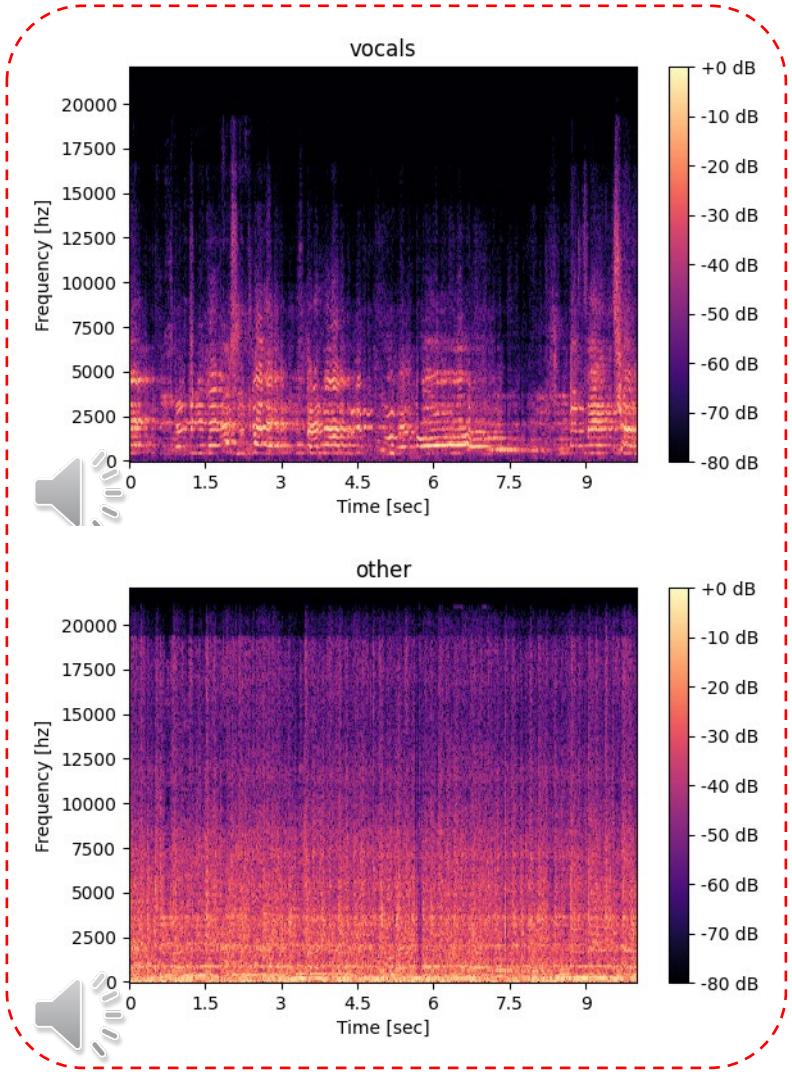


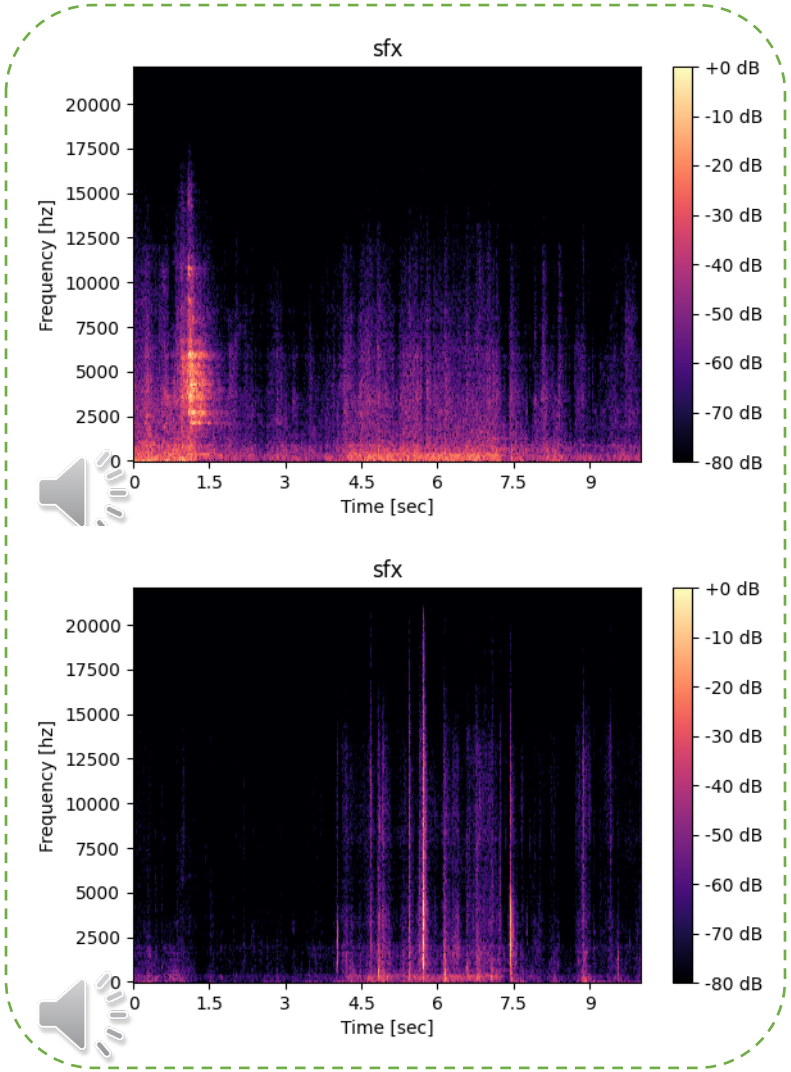


TUSS output with prompts: <Speech>, <Vocals>, <Other inst.>, <SFX>, <SFX>









Conclusions on TUSS

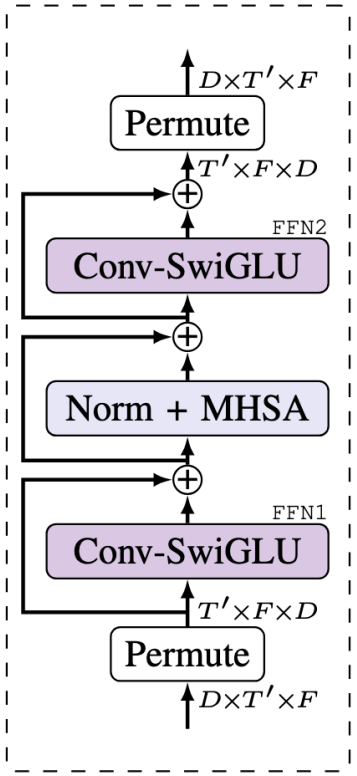
- First model that can truly tackle general source separation at the level of specialist models
- Generalizes to new tasks unseen during training

- TUSS is great, but there are practical shortcomings:
  - Offline, i.e., non-causal
  - Large compute requirements even though parameter count is small relative to LLM

- Only single-channel so far: can we be flexible in the number of channels too?

- Can we apply the task-aware idea to other types of models?

What is the main compute bottleneck?



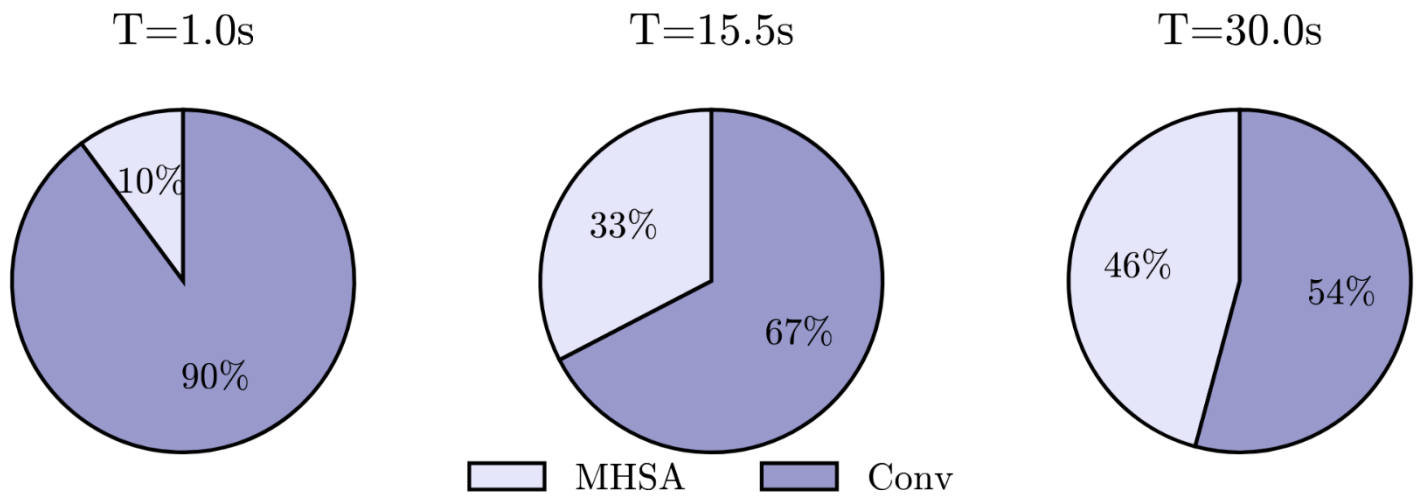


Fig. 2: Compute breakdown of MHSA vs. convolutions for different chunk lengths.

- Typical chunk sizes for source separation networks are between 2s – 12s
- We measure inference speed in terms of multiply accumulate (MAC) operations
- The convolution layers are more expensive than multi-head self attention (MHSA)

How to go FasTUSS? [Paissan+, WASPAA2025]

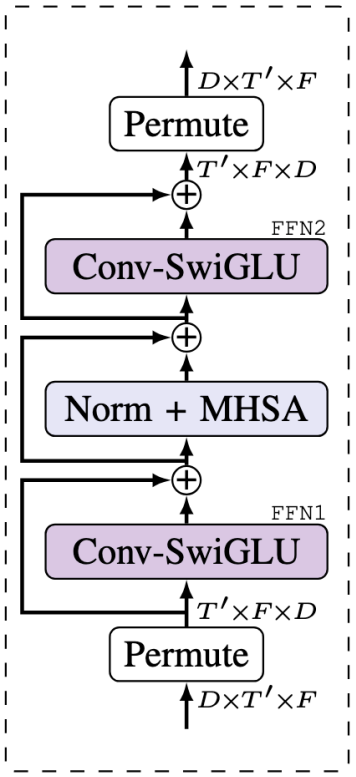


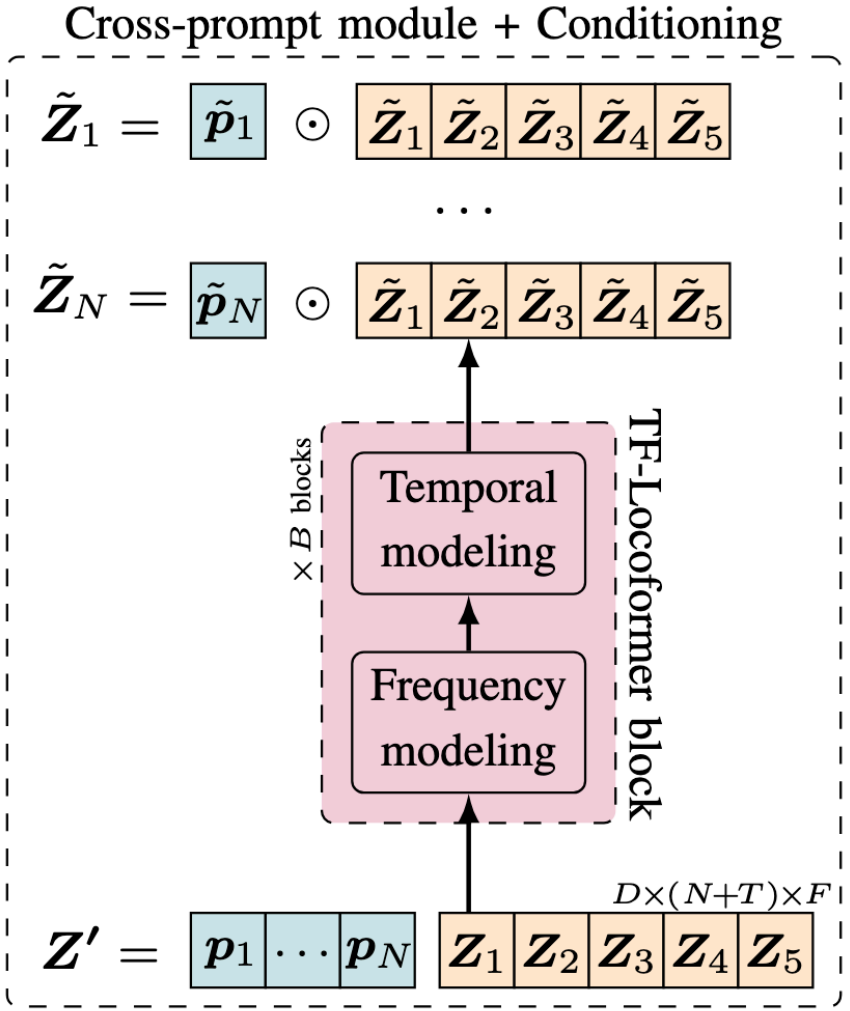
Table 1: Comparison of various speedup and conditioning configurations.

ID	S	G	FFN1	Params (M)	MAC (G)	Δ SNR [dB]
1	1	1	✓	11.1	43.1	0.0
2	2	1	✓	11.1	26.2	−0.2
3	4	1	✓	11.1	17.7	−0.5
4	1	8	✓	10.8	40.5	−1.6
5	1	1	✗	8.9	24.4	−0.3
6	2	1	✗	8.9	16.0	−0.6
7	4	1	✗	8.9	11.7	−0.4
8	4	8	✗	7.5	8.3	−1.2

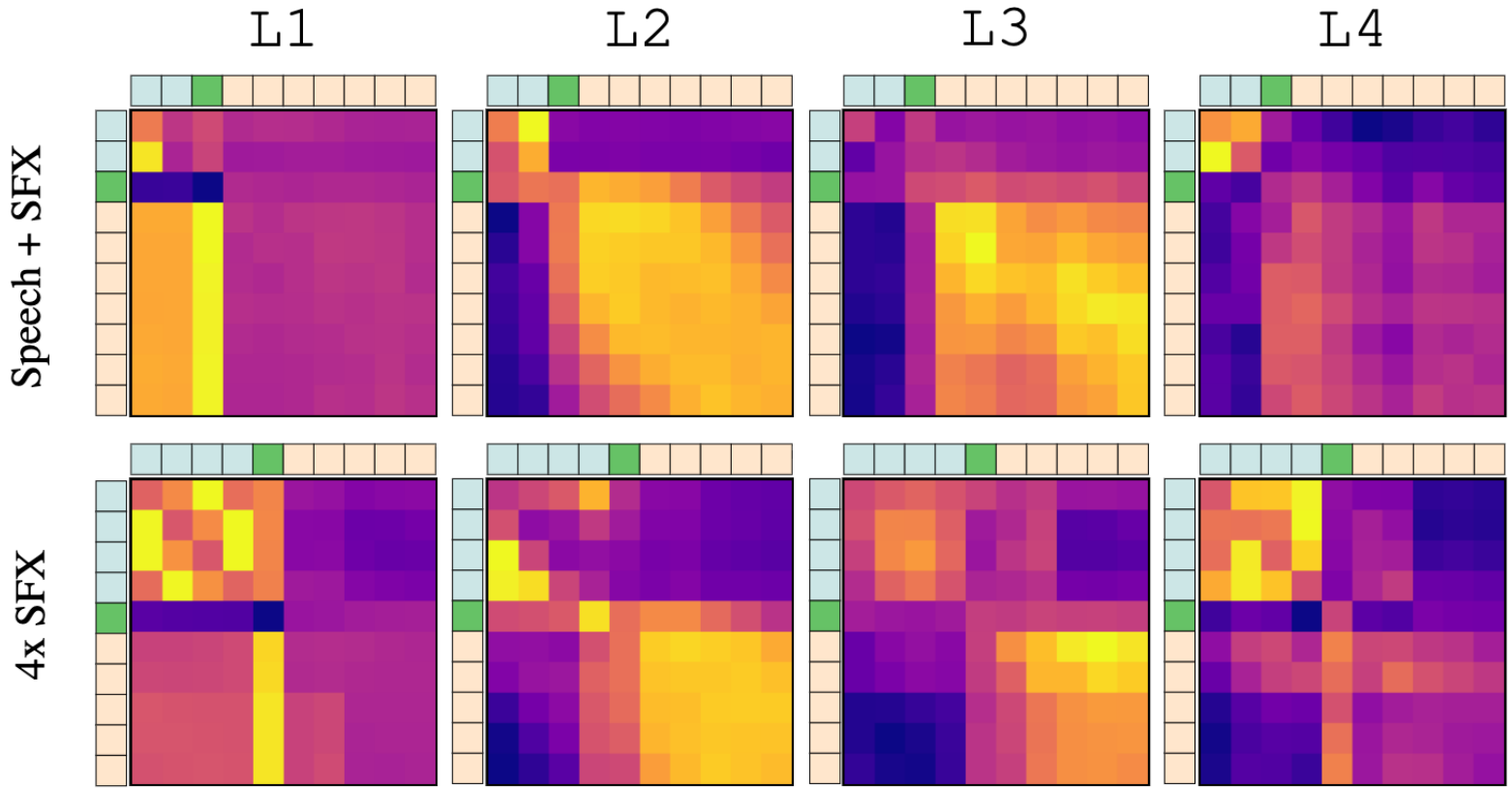
- Benchmark on suite of datasets: VCTK-DEMAND, WHAM!, FUSS, MUSDB-HQ, DnR
- Removing first FFN block achieves comparable performance with considerable drop in compute (~43%)
- We increase the stride in the convolutions with negligible performance drop
- We define **FasTUSS-11.7G** as configuration ID7 and **FasTUSS-8.3G** as configuration ID8 from Table 1, as both are good options for maintaining strong performance while reducing compute
  - reduce TUSS operations by 73% (0.4 dB performance drop) or 81% (1.2 dB performance drop)

Causal TUSS

- Straightforward to make convolutions causal
- MHSA inside cross-prompt module needs special care
  - mixture processing must be causal
  - prompts must be updated in MHSA without leaking information during training stage
  - using a standard attention mask on the prompts would lead to the first prompt (first element of Z') never being updated with information from other prompts or the mixture.

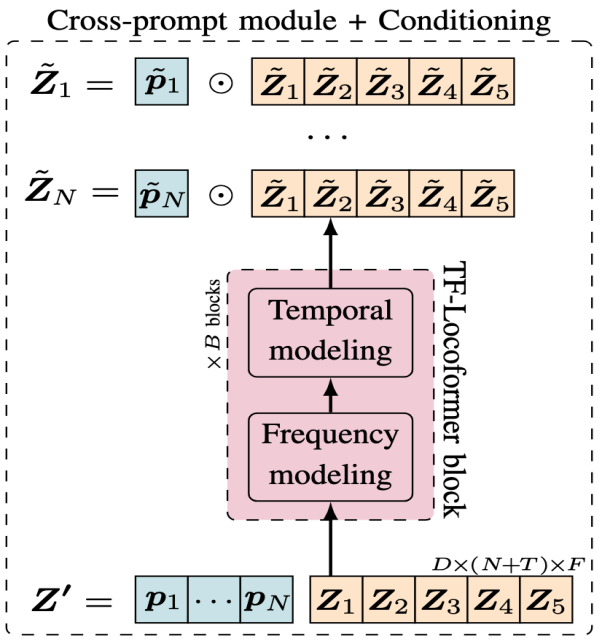


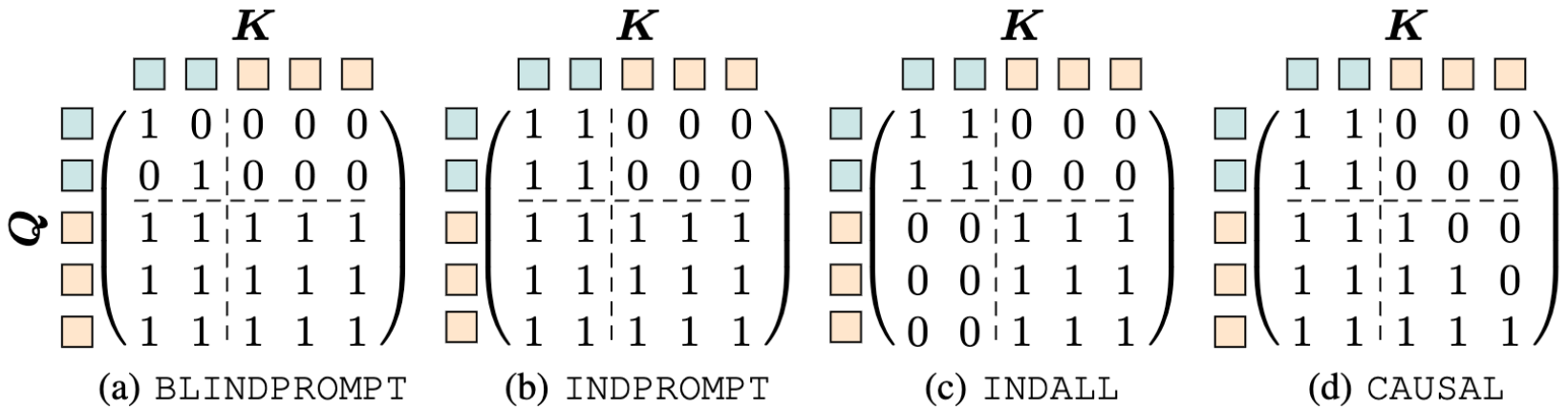
Prompts and mixtures cross-condition each other in TUSS



Attention maps in log scale. Teal, green, and orange tokens represent the prompts, the <SOS> token, and the mixture, respectively. L* represent the layers on the temporal modeling path of the cross-prompt module.

Impact of prompt-to-mixture and mixture-to-prompt conditioning



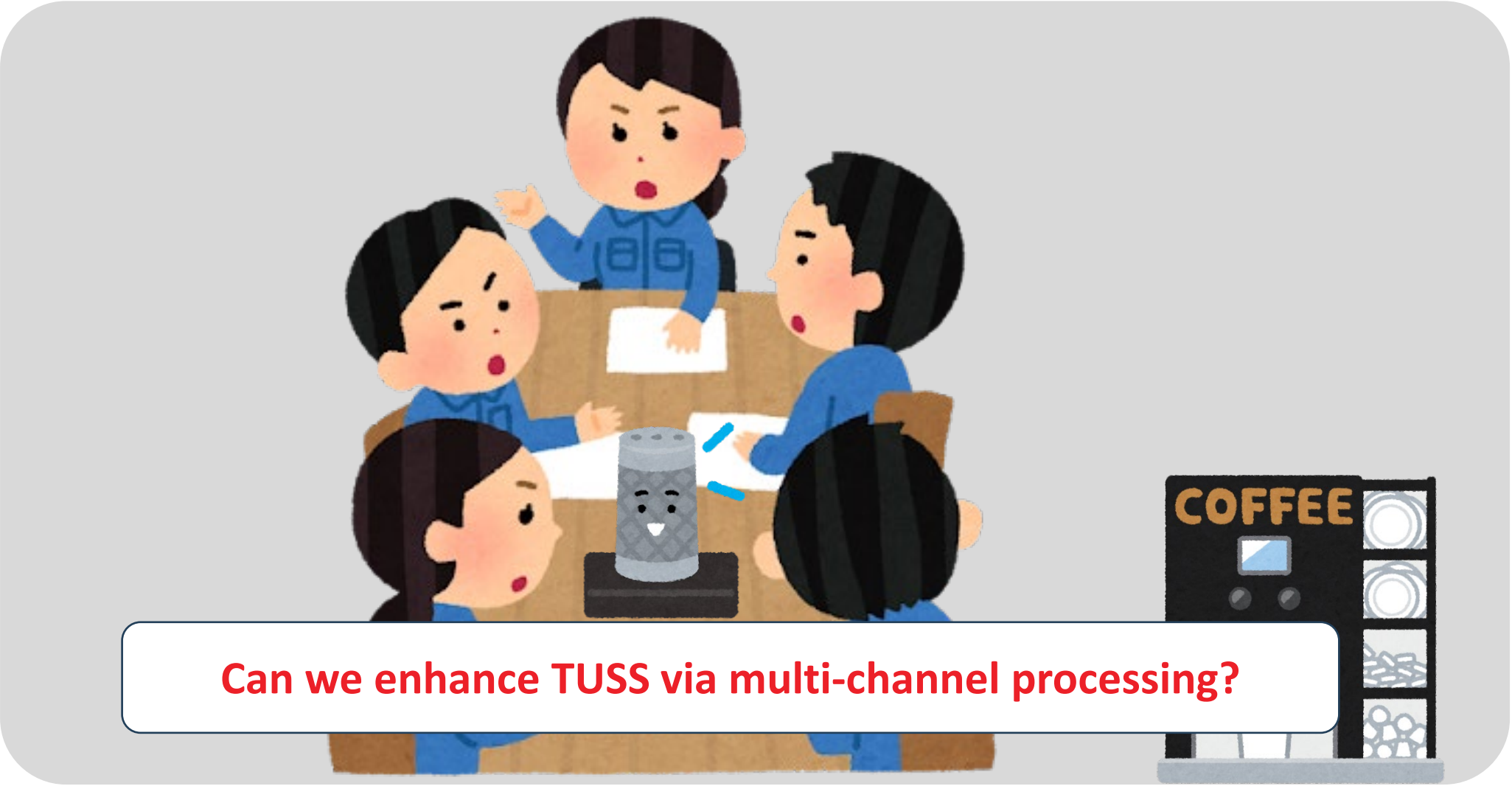


ID	S	G	FFN1	Params (M)	MAC (G)	Δ SNR [dB]
BLINDPROMPT	1	1	✓	11.1	43.1	−1.3
INDPROMPT	1	1	✓	11.1	43.1	−0.2
INDALL	1	1	✓	11.1	43.1	−3.2
CAUSAL	1	1	✓	11.1	43.1	−1.8

- Especially important for prompts to be conditioned on each other, to handle repeated prompts and for context definition, performance significantly drops otherwise (BLINDPROMPT)
- Significant drop in performance when the mixture processing is not conditioned on the prompts and vice-versa (INDALL)
- We derive a **causal version of TUSS**, where prompts are not conditioned on the mixture, but still conditioned on the context (CAUSAL)

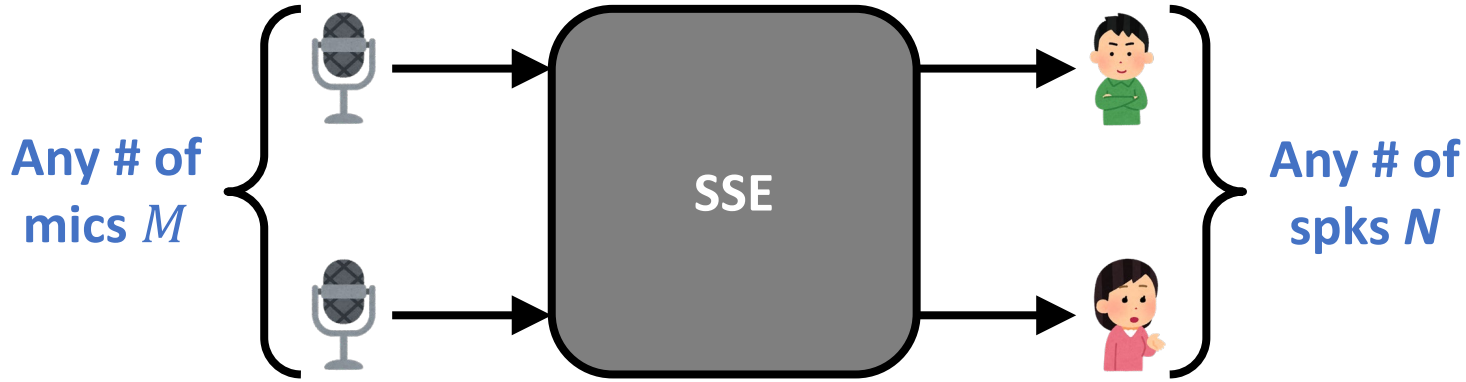
Speech Separation and Enhancement (SSE) for Various Situations

- Prompt-conditional separation allows specifying the number of speakers during inference.
  → **TUSS can deal with audio mixtures with an arbitrary number of speakers.**



Flexible and Versatile SSE

- We aim to realize an SSE system that **can handle any number of microphones and speakers.**



- **Existing systems are still limited** in some aspects of flexibility and/or controllability.
  - Controllability refers to whether users can specify the number of speakers during inference.

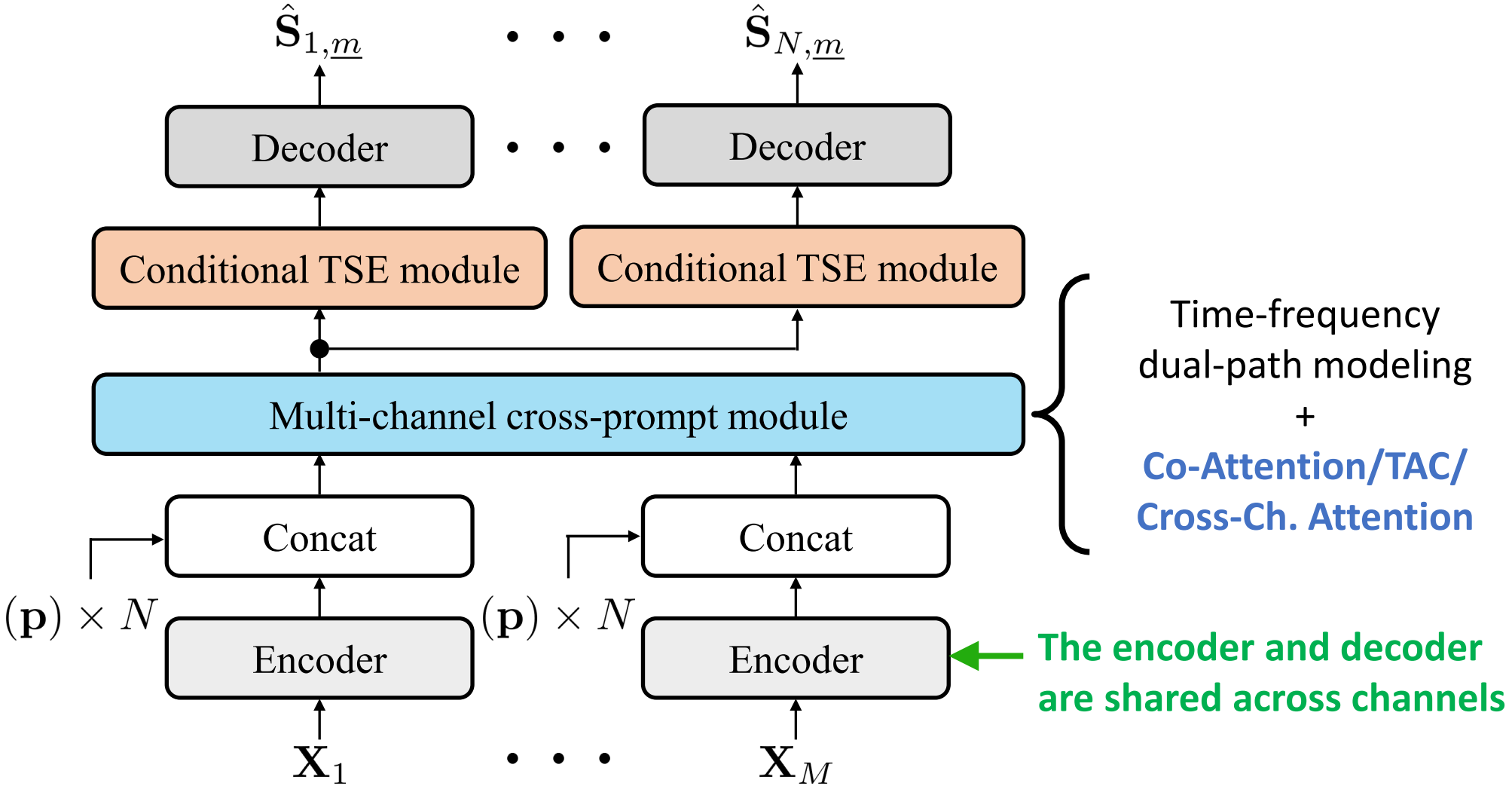
	Flexible M	Flexible N	Controllability
TPARN [25], USES [26]	✓	✗	✗
SepEDA [15], TUSS [18]	✗	✓	✓
VarArray [27]	✓	✓	✗
DNN-IVA [28]	✓†	✓	✓
FlexIO (Proposed)	✓	✓	✓

†DNN-IVA is not applicable to single-channel scenarios

FlexIO: An Array-Agnostic Extension of TUSS

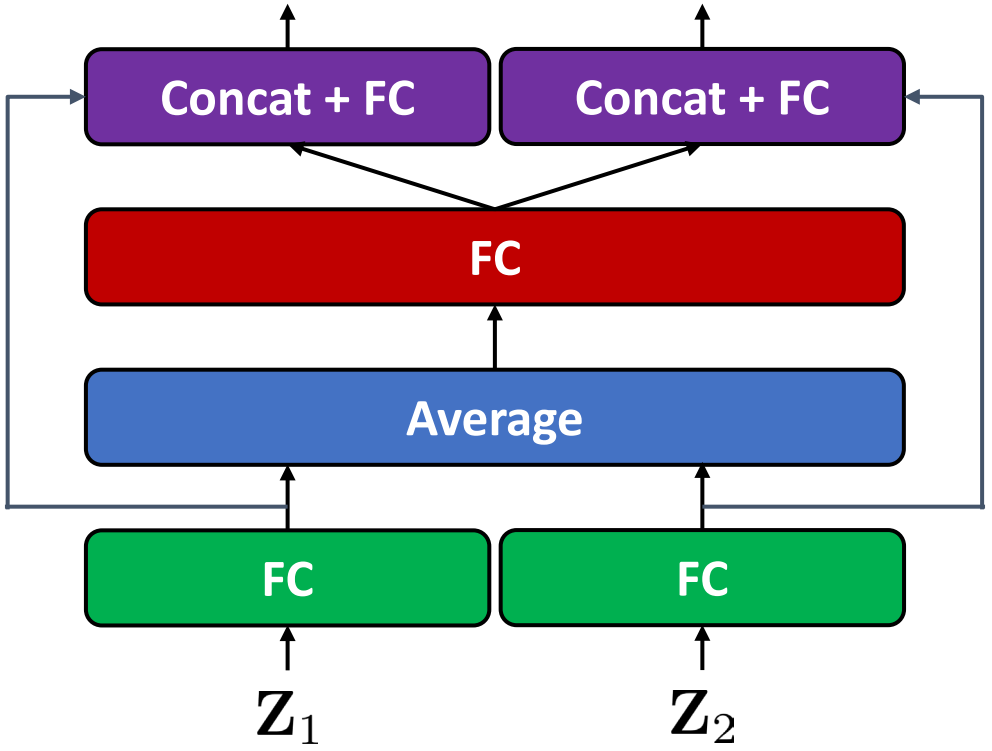
[Masuyama+, ICASSP2026]

- We incorporate **array-agnostic channel communication** into the cross-prompt module.
- FlexIO performs conditional TSE on the representation from the reference channel.



Transform Average Concatenate (TAC)

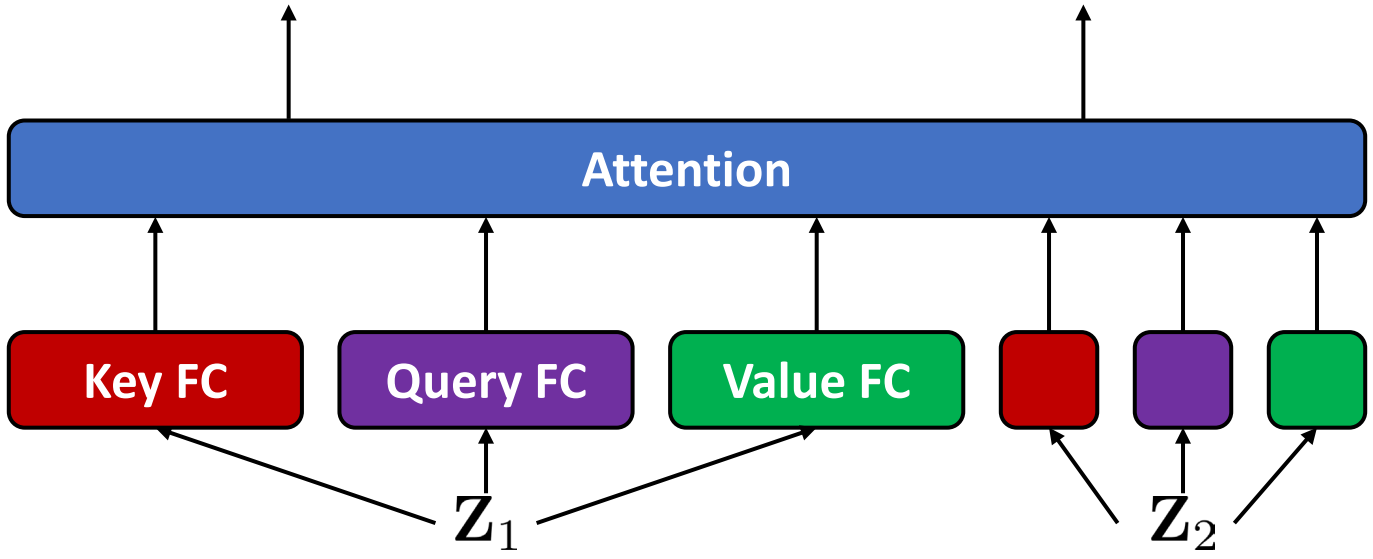
- TAC aggregates the channel-wise representation by **average pooling.**



Skip connection from the input to the output is omitted.

Cross-Channel Attention

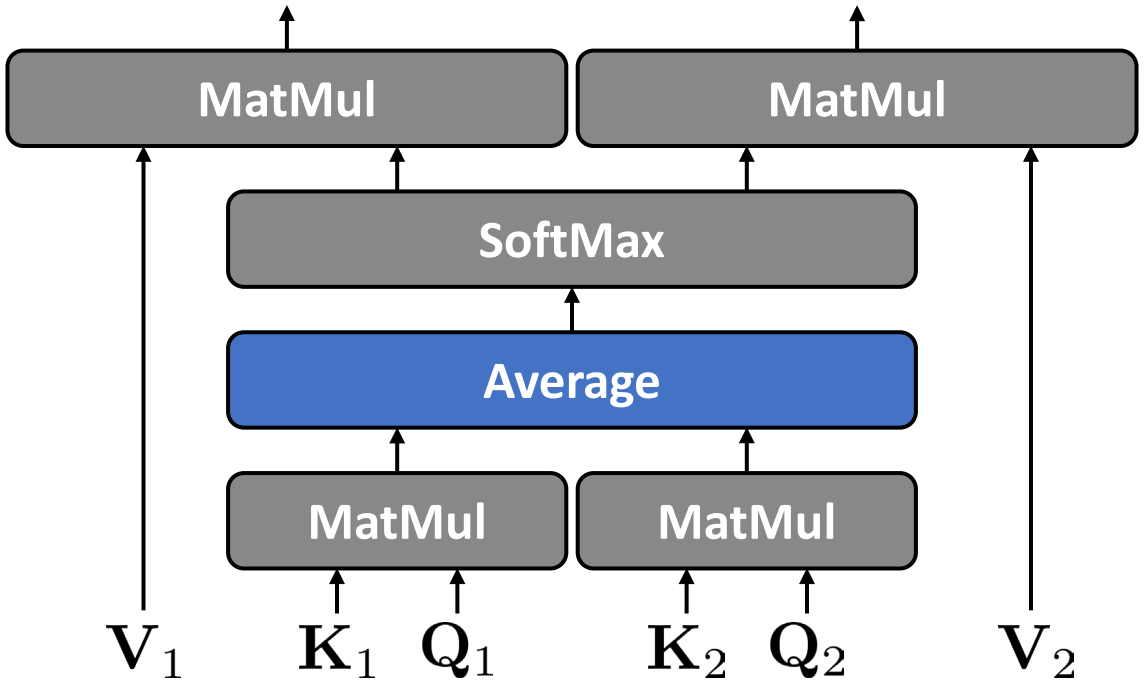
- Cross-channel attention dynamically exchanges information across channels via **attention.**



Skip connection from the input to the output is omitted.

Co-Attention

- Co-attention **shares the same attention matrix across channels** within dual-path modeling.
  - This approach does not require additional network parameters.



Experimental validation

- Train on several enhancement and separation datasets
  - with 1, 2, or 4 microphones
  - with 1, 2, or 3 speakers
- 3-channel and 5-channel settings never seen during training to assess generalization

Table 2. Datasets used in our SSE experiments. "A" and "R" in the condition denote anechoic and reverberant settings. The 3- and 5-channel data are excluded from the training and validation sets.

	#Ch M	#Spks N	Acoustic condition
CHiME-4 [34]	$\{1, 2, 4\}, \{3, 5\}$	1	Noisy A
WSJ0-mix [6]	1	$\{2, 3\}$	Clean
WHAM! [35]	1	$\{1, 2\}$	Noisy A
WHAMR! [36]	$\{1, 2\}$	$\{1, 2\}$	Noisy A/R
WSJ1-CHiME [28]	$\{2, 4\}, 3$	$\{2, 3\}$	Noisy R

Evaluation on Speech Enhancement

- **FlexIO consistently improves the enhancement performance over single-channel TUSS** when the number of microphones is more than 1.
- Among medium-size FlexIO,
  - cross-channel attention performs slightly better with up to 2 channels.
  - co-attention outperforms the other modules on the CHiME-4 dataset.
- **Large FlexIO outperforms an existing array-agnostic speech enhancement model (USES).**

Table 3. Speech enhancement performance under diverse conditions. The numbers in parentheses denote the numbers of speakers and microphones, i.e., (N-M). For the WHAMR! dataset, "A" and "R" indicate the anechoic and reverberant conditions, respectively. "Comm." shows the channel communication mechanism, where "1ch" indicates that speech enhancement is performed solely on the reference channel. The best and second best results are highlighted in blue and orange, respectively

			WHAM! (1-1)			WHAMR! A (1-2)			WHAMR! R (1-2)			CHiME-4 (1-4)			CHiME-4 (1-5)		
	Comm.	#Params (10^6)	SDR	STOI	PESQ	SDR	STOI	PESQ	SDR	STOI	PESQ	SDR	STOI	PESQ	SDR	STOI	PESQ
USES [26]	TAC	3.05	10.2	85.7	1.65	15.8	96.4	2.55	13.8	96.0	2.51	18.3	96.6	2.46	18.3	97.8	2.95
TUSS [18]	1ch	3.42	13.6	94.2	2.29	13.6	94.1	2.28	12.4	93.7	2.24	17.1	95.7	2.31	17.1	95.7	2.31
FlexIO	TAC (M)	3.59	13.5	94.0	2.25	15.4	95.6	2.55	14.2	95.3	2.50	19.3	97.3	2.54	19.6	97.5	2.60
	ChAtt (M)	3.49	13.6	94.1	2.25	15.6	95.9	2.57	14.8	95.7	2.56	19.5	97.4	2.55	20.2	97.6	2.61
	CoAtt (M)	3.42	13.5	94.0	2.26	15.5	95.7	2.56	14.4	95.4	2.52	20.8	97.8	2.72	21.7	98.1	2.81
	CoAtt (L)	7.35	13.8	94.4	2.31	15.8	96.0	2.61	14.9	95.8	2.58	21.3	98.0	2.86	22.3	98.3	2.96

Evaluation on CHiME-4 Real Recordings

- We evaluate the generalization capability of FlexIO on real multi-channel recordings.
- **USES deteriorates both DNSMOS and WER on the real 5-channel recordings.**
  - Its extension, U2-C, improves DNSMOS but still **degrades WER from the noisy mixture.**
- **FlexIO achieves promising DNSMOS,** and **WER is improved** as well as MVDR beamforming.

Table 5. DNSMOS and recognition performance on CHiME-4 recordings. FlexIO adopts co-attention for channel communication.

	Simulated				Real			
	DNSMOS		WER (%)		DNSMOS		WER (%)	
	1ch	5ch	1ch	5ch	1ch	5ch	1ch	5ch
Noisy/MVDR	2.08	2.57	5.8	4.0	1.46	1.95	6.7	4.5
USES [26]	3.03	3.22	15.2	4.4	3.07	1.58	7.4	85.9
U2-C [41]	-	-	-	-	-	3.08	-	10.3
FlexIO (M)	3.11	3.17	6.0	3.9	2.91	3.05	6.8	4.5

Evaluation on Speech Separation

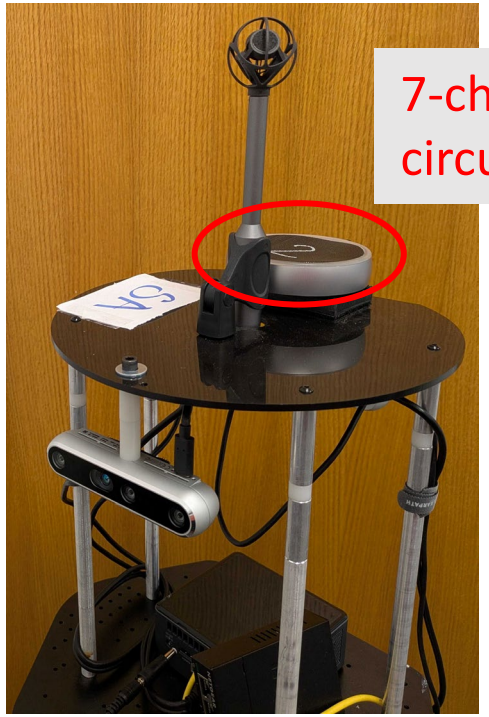
- DNN-IVA is not applicable to single-channel scenarios as it relies on array signal processing
- Large FlexIO outperforms the situation-specific TF-GridNet, **showing the efficacy of FlexIO**
- **Similarly to the enhancement experiments, FlexIO outperforms single-channel TUSS**

Table 4. Speech separation performance in various settings. FlexIO and TUSS are the same models as those used for speech enhancement in Table 3. While DNN-IVA uses the same model for all conditions, TF-GridNet† is separately trained for the 1- and 2-channel conditions.

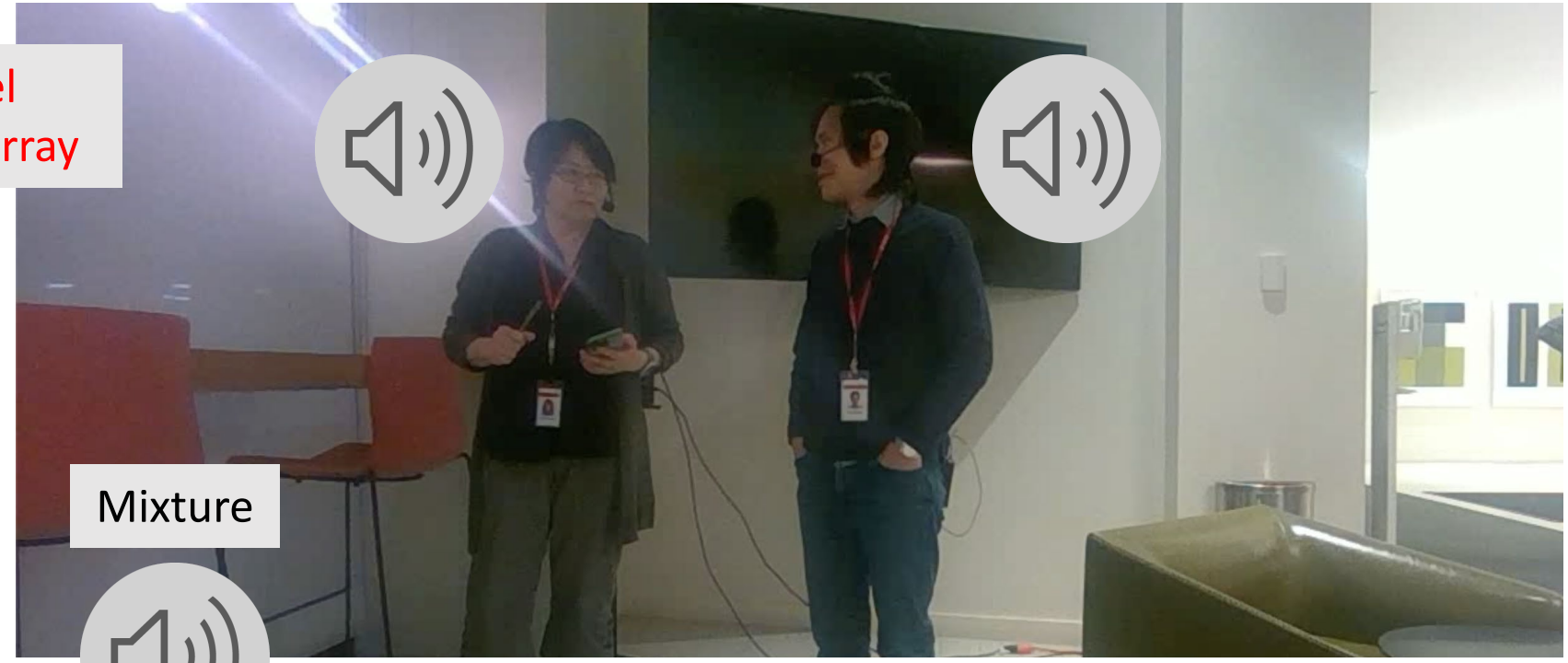
			WHAMR! R (2-1)			WHAMR! R (2-2)			WSJ1-CHiME (2-2)			WSJ1-CHiME (2-4)			WSJ1-CHiME (3-3)		
	Comm.	#Params (10^6)	SDR	SIR	PESQ	SDR	SIR	PESQ	SDR	SIR	PESQ	SDR	SIR	PESQ	SDR	SIR	PESQ
DNN-IVA [28]	-	5.13	N/A	N/A	N/A	-	-	-	10.7	24.1	-	-	-	-	7.7	20.1	-
TF-GridNet† [5]	-	8.38	9.3	27.0	1.79	11.7	29.6	2.20	-	-	-	-	-	-	-	-	-
TUSS [18]	1ch	3.42	9.5	25.3	1.82	9.5	25.3	1.82	14.9	29.1	2.73	15.1	29.4	2.75	11.4	23.1	2.15
FlexIO	TAC (M)	3.59	8.9	24.0	1.77	11.8	28.5	2.14	18.5	33.9	3.22	19.5	25.0	3.37	15.7	28.4	2.81
	ChAtt (M)	3.49	9.0	23.7	1.75	12.4	29.5	2.22	19.5	34.9	3.34	20.6	36.2	3.49	16.9	30.1	2.97
	CoAtt (M)	3.42	9.1	24.4	1.78	12.1	29.0	2.18	18.9	34.7	3.27	20.6	36.4	3.49	16.7	30.0	2.96
	CoAtt (L)	7.35	9.7	25.5	1.84	12.5	29.6	2.22	19.6	35.0	3.36	21.6	37.3	3.60	17.3	30.4	3.03

Evaluation on Unseen In-House Recordings

- FlexIO is trained on **simulated mixtures up to 4 channels**
- **Still performs well on in-house 7-channel recordings** for human-computer interaction



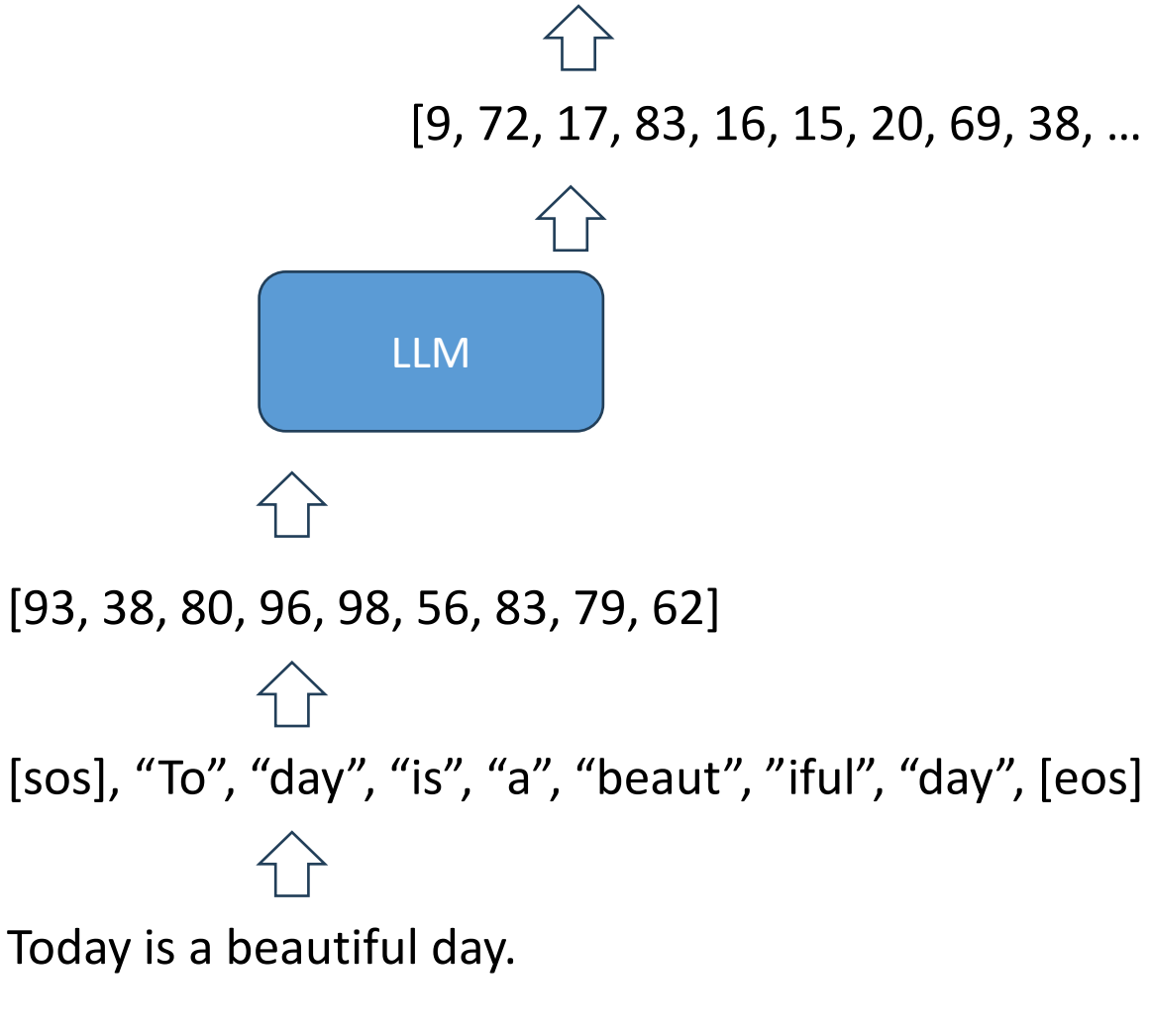
7-channel
circular array



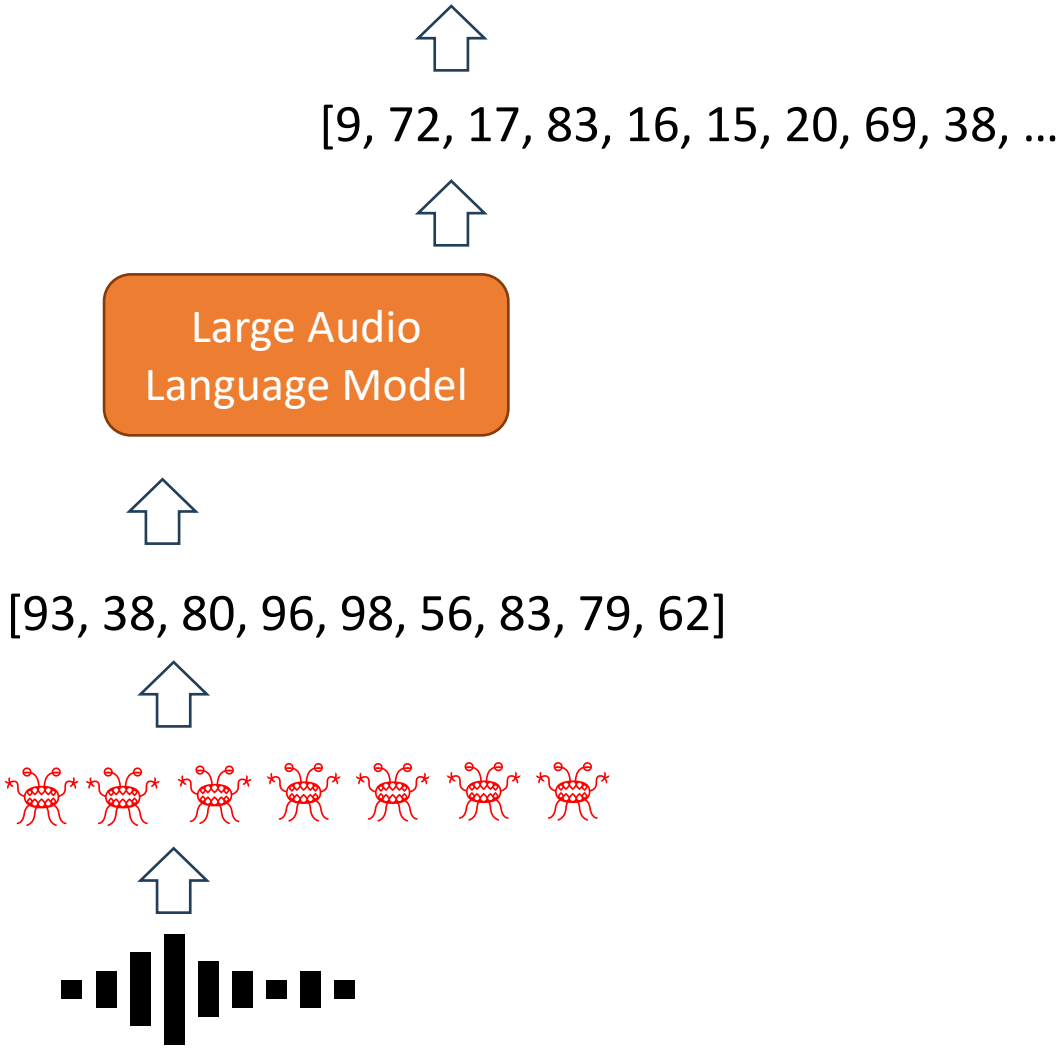
Mixture

Source-awareness in neural audio codecs?

- Large Language Model (LLM)

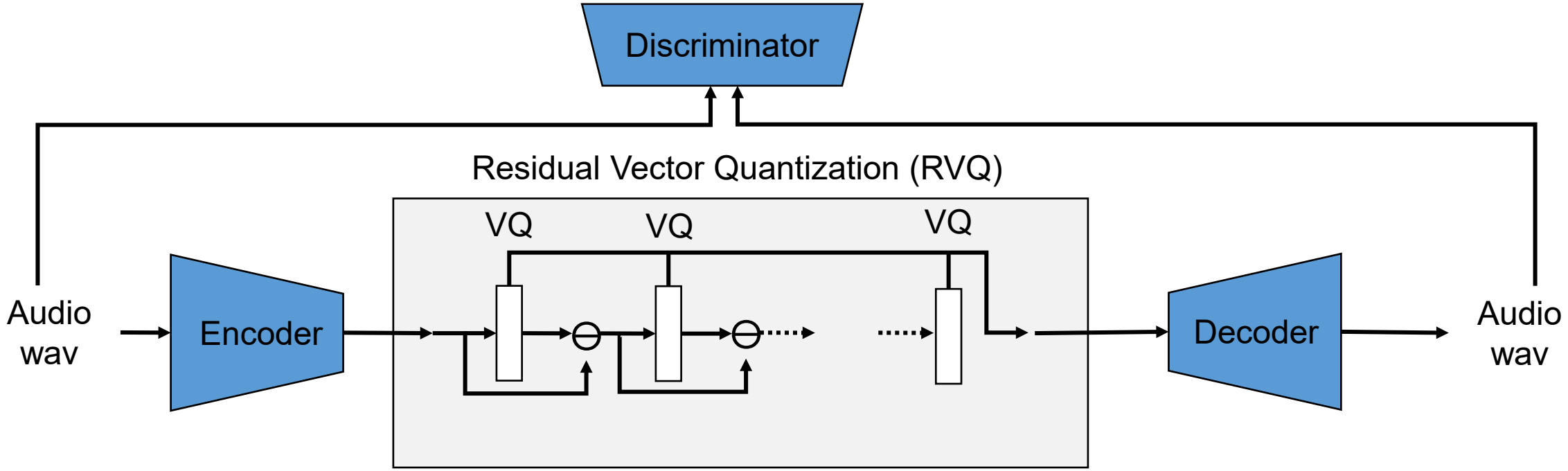


- Large Audio Language Model

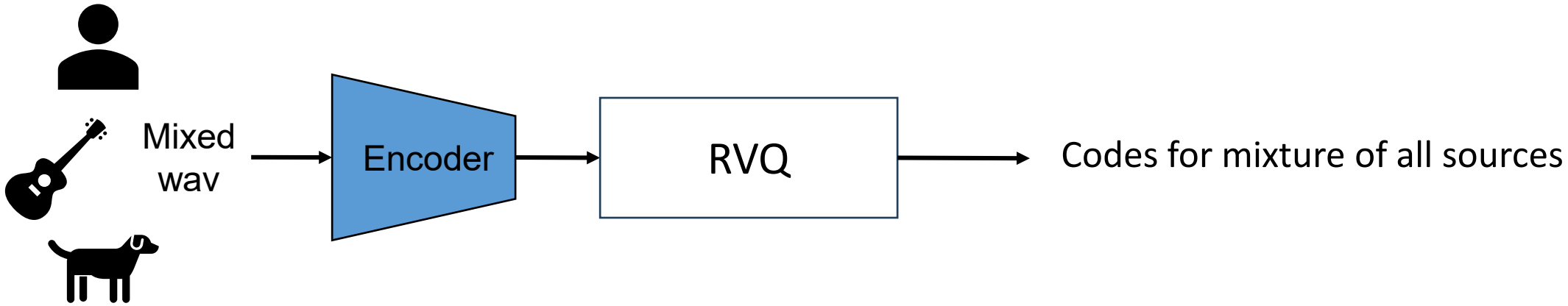


Conventional neural audio codecs

- Typical architecture of NACs
  - Encoder estimates downsampled and continuous representation
  - RVQ estimates discrete codes from the latent representation
  - Decoder estimates the waveform from the discrete codes
  - Trained using Generative Adversarial Networks (GAN) framework

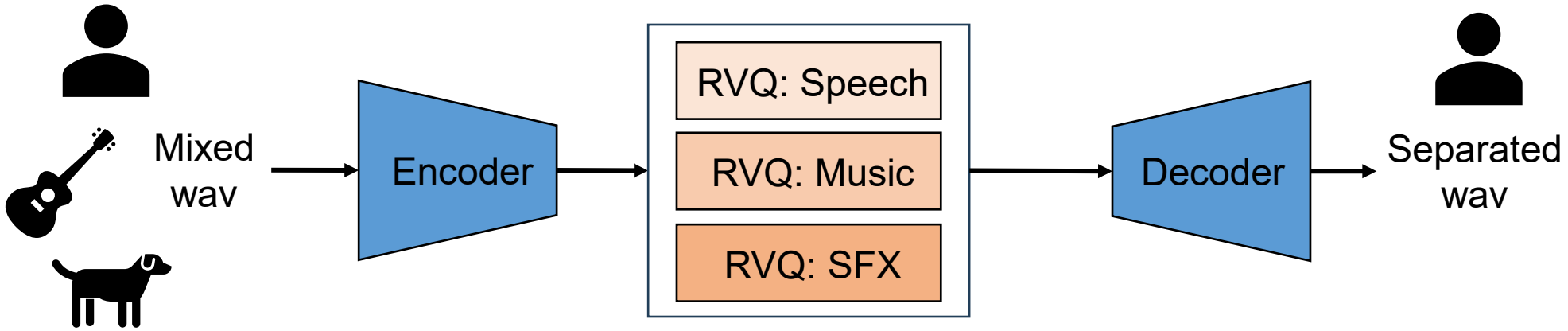


No disentanglement of acoustic information



Can we disentangle the sources at the token level
and make the Audio LM's life easier?

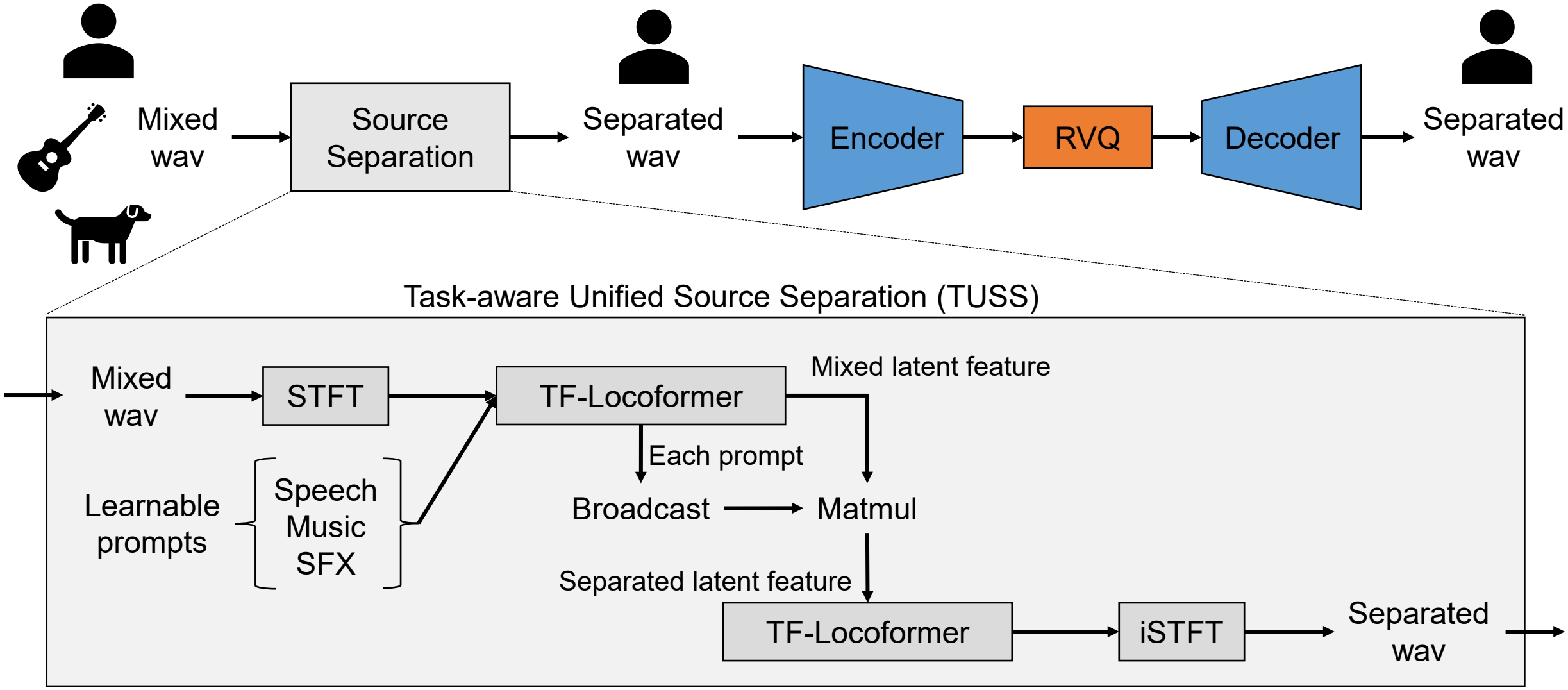
Source Disentangled Codec (SDCodec) [Bie et. al., ICASSP 2025]



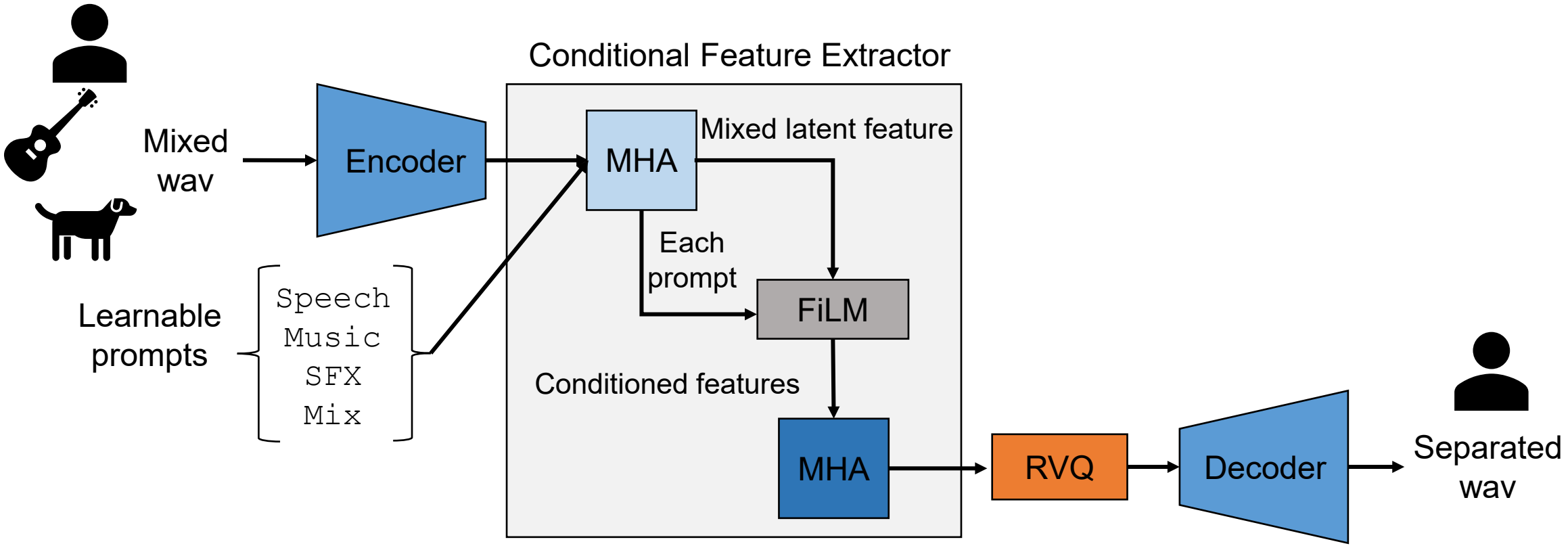
- Separation relies only on RVQs, which are specific to each source domain
- Unable to separate mixtures of sources from the same domain (e.g., two speakers)

NACs with source separation: Cascaded model

- Apply source separation before NACs



SUNAC: Source-aware Unified Neural Audio Codec [Aihara+, ICASSP2026]



- Add Multi-Head Attention (MHA) and Feature-wise Linear Modulation (FiLM) before RVQ
- Enables separation of mixtures within the same source domain (e.g., two speakers)

Experimental conditions

- Training data for NACs
  - Speech: 2600 hours
    - DNS 5 speech, MUSAN-speech
  - Music: 3800 hours
    - MTG-Jamendo, MUSAN-music
  - SFX: 260 hour
    - DNS5-sfx, MUSAN-sfx, WHAM!
  - Sample 1-3 sources
    - up to 2x `<Speech>`
    - no repeat of `<Music>` and `<SFX>`
- Valid / Test data for NACs
  - Valid / Test division from DnR dataset
- Metrics for evaluation
  - Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [dB]
  - ViSQOL (perceptual auto quality assessment)
  - Higher is better on both metrics

Comparison of model parameters and computational cost (MAC) measured per 1.0 second in GMACs

Method	Params (M)	Const [G]	Per source [G]
TUSS	11.1	21.1	10.5
FasTUSS	11.1	4.1	2.1
DAC	74.1		41.0
DACT	66.4		12.9
TUSS-DAC	85.2	21.1	51.5
FasTUSS-DACT	77.5	4.1	14.9
SDCodec	74.8	12.6	28.4
SDCodecT	67.1	3.9	9.0
SUNAC	69.2	3.5	9.5

Experimental results

- Separation from {<Speech>,<Music>,<SFX>}

Table 3. Reconstruction results for the mixture and separated sources from {<Speech>, <Music>, <SFX>} (no repeated prompt).

	Mix		Speech		Music		SFX	
Model	**SI-SDR ↑**	**VisQOL ↑**	**SI-SDR ↑**	**VisQOL ↑**	**SI-SDR ↑**	**VisQOL ↑**	**SI-SDR ↑**	**VisQOL ↑**
TUSS-DAC	–	–	13.07 ± 2.46	3.72 ± 0.38	4.70 ± 4.18	4.25 ± 0.18	4.82 ± 5.17	4.19 ± 0.18
FasTUSS-DACT	–	–	12.29 ± 2.50	3.53 ± 0.39	3.92 ± 4.18	4.26 ± 0.16	3.80 ± 5.38	4.17 ± 0.17
SDCodec†	6.39 ± 3.19	4.52 ± 0.05	10.78 ± 2.99	3.50 ± 0.43	1.74 ± 3.57	4.26 ± 0.13	2.17 ± 4.33	4.20 ± 0.15
SDCodec	8.24 ± 2.83	4.56 ± 0.05	11.40 ± 3.08	3.64 ± 0.41	1.21 ± 3.58	4.10 ± 0.21	1.26 ± 4.27	4.18 ± 0.17
SDCodecT	7.30 ± 3.06	4.54 ± 0.06	11.32 ± 2.89	3.64 ± 0.37	1.75 ± 4.45	4.09 ± 0.22	1.42 ± 4.76	4.09 ± 0.22
SUNAC	6.48 ± 3.11	4.52 ± 0.06	11.56 ± 3.00	3.68 ± 0.36	1.98 ± 4.62	4.14 ± 0.20	2.10 ± 4.94	4.16 ± 0.19

  - Comparable performance to SDCodec
  - Codec based methods achieve lower SI-SDR than FasTUSS-DACT (phase issue?) but comparable VisQOL, at lower complexity

Experimental results

- Separation from repeated prompts
  - SDCodec cannot handle this case
  - SUNAC comparable with cascaded approaches, even on 4-prompt case for which it wasn't trained

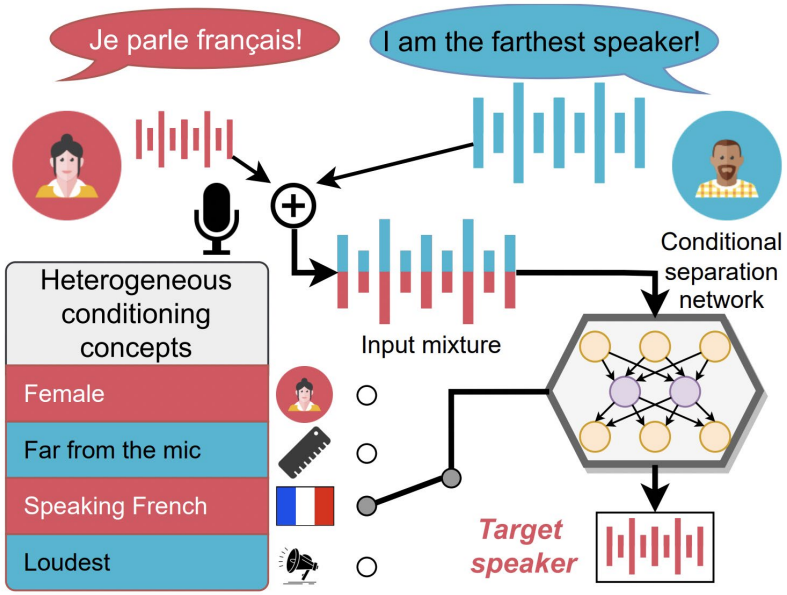
Table 4. Separated results from {<Speech>,<Speech>}.

Model	SI-SDR ↑	VisQOL ↑
TUSS-DAC	13.35 ± 3.80	4.08 ± 0.39
FasTUSS-DACT	10.73 ± 4.66	3.83 ± 0.46
SDCodec†	0.00 ± 2.83	3.04 ± 0.61
SDCodec	0.00 ± 2.83	3.04 ± 0.62
SDCodecT	0.00 ± 2.83	3.09 ± 0.59
SUNAC	11.80 ± 3.07	4.12 ± 0.42

Table 5. Reconstruction results for mixed source and separated results from {<Speech>, <Speech>, <Music>, <SFX>}.

	Mix		Speech		Music		SFX	
Model	SI-SDR ↑	VisQOL ↑	SI-SDR ↑	VisQOL ↑	SI-SDR ↑	VisQOL ↑	SI-SDR ↑	VisQOL ↑
TUSS-DAC	–	–	9.07 ± 3.38	3.40 ± 0.47	2.75 ± 3.96	4.20 ± 0.17	3.05 ± 5.23	4.13 ± 0.18
FasTUSS-DACT	–	–	6.98 ± 3.92	3.08 ± 0.38	2.09 ± 3.82	4.21 ± 0.16	2.07 ± 5.37	4.11 ± 0.18
SDCodec†	6.55 ± 2.59	4.49 ± 0.06	-0.95 ± 3.29	2.58 ± 0.53	-0.69 ± 3.64	4.20 ± 0.13	-0.15 ± 4.69	4.15 ± 0.15
SDCodec	8.39 ± 2.31	4.54 ± 0.05	-1.00 ± 3.34	2.64 ± 0.54	-1.62 ± 3.77	4.07 ± 0.21	-0.96 ± 4.44	4.12 ± 0.17
SDCodecT	7.45 ± 2.52	4.51 ± 0.06	-0.95 ± 3.58	2.60 ± 0.56	-1.15 ± 4.45	4.07 ± 0.22	-0.61 ± 4.81	4.11 ± 0.17
SUNAC	6.38 ± 2.54	4.51 ± 0.06	7.46 ± 3.41	3.33 ± 0.45	0.15 ± 4.29	4.11 ± 0.20	0.25 ± 4.97	4.11 ± 0.19

Looking forward



- Expanding to richer prompts and conditions
  - Semantic / spatial prompts (roles, attributes, relationships: “main speaker”, “speaker on the left”)
  - Grounding via multimodal prompts (text + audio cues: “the speaker that sounds like this”)
  - Beyond additive distortions (clipping, codec, …)

[Tzinis+, Interspeech2022]

- Weakly-supervised/unsupervised approaches for large-scale training
  - Self-remixing [Saijo and Ogawa, ICASSP2023]
  - Noisy CLAP training for text-queried target sound extraction [Saijo+, ICASSP2025]

- Generative models for separation

- Integration with Audio/Speech LMs

MITSUBISHI
ELECTRIC

Changes for the Better

MITSUBISHI ELECTRIC RESEARCH LABORATORIES