A *Sidecar* Separator Can Convert a Single-Talker Speech Recognition System to a Multi-Talker One

Lingwei Meng, Jiawen Kang, Mingyu Cui, Yuejiao Wang,
Xixin Wu, Helen Meng

Human-Computer Communications Laboratory,
The Chinese University of Hong Kong

Image source: https://ridermagazine.com/2011/05/13/a-short-history-of-sidecars/

Outline

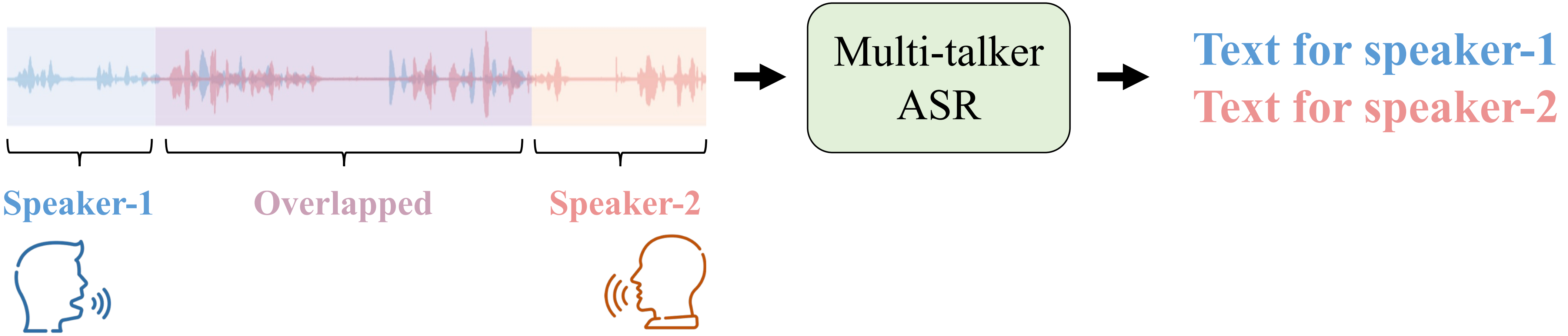
1. **Background**
2. Objective
3. Proposed Approach
4. Experiments
5. Conclusion

Image source: https://ridermagazine.com/2011/05/13/a-short-history-of-sidecars/

1. Background

Definition of Multi-talker Speech Recognition:

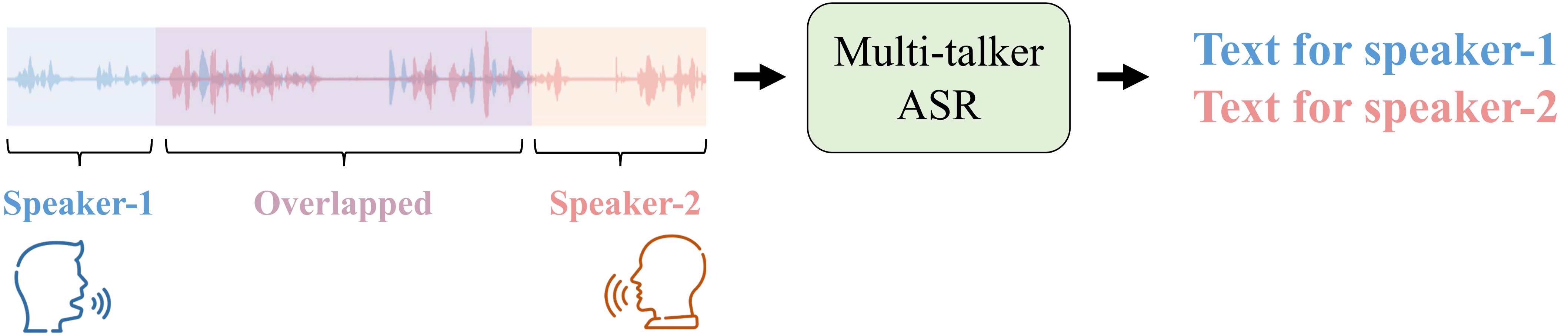
To transcribe texts for different speakers from multi-talker overlapped speech



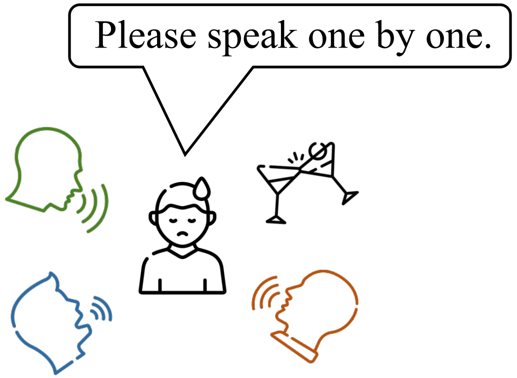
1. Background

Definition of Multi-talker Speech Recognition:

To transcribe texts for different speakers from multi-talker overlapped speech



It remains a significant challenge!



1. Background – Literature Review

Existing multi-talker ASR strategies have their drawbacks:

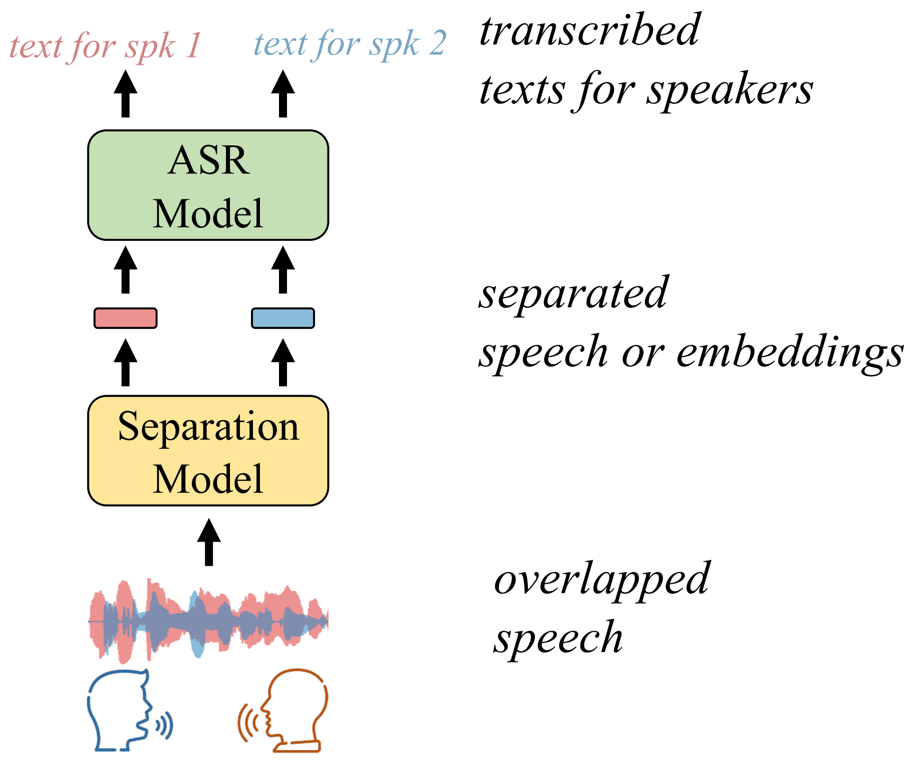
1. Background – Literature Review

Existing multi-talker ASR strategies have their drawbacks:

Existing strategy I:

Cascade architecture of Separation and ASR

- Need further joint fine-tuning
- The fine-tuned modules cannot work well individually anymore.



[1] Shane Settle et al. “End-to-End Multi-Speaker Speech Recognition,” ICASSP 2018
[2] Song Li et al. "Real-time End-to-End Monaural Multi-speaker Speech Recognition," Interspeech 2021

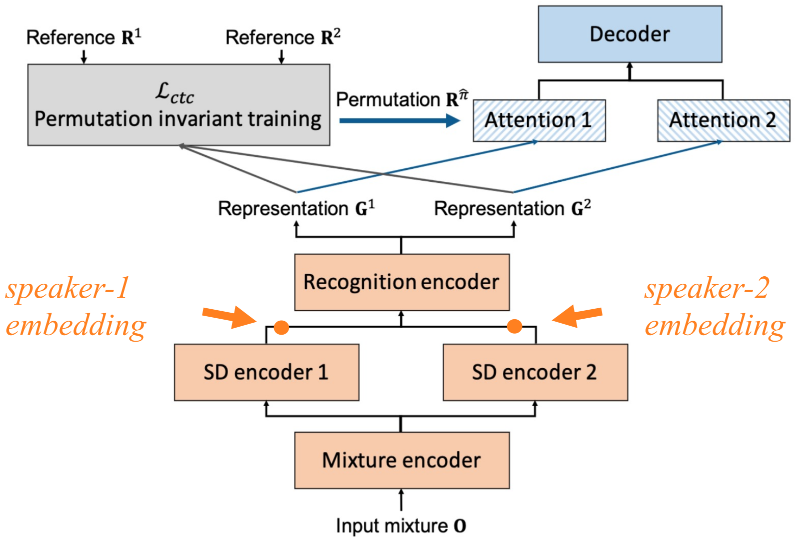
1. Background – Literature Review

Existing multi-talker ASR strategies have their drawbacks:

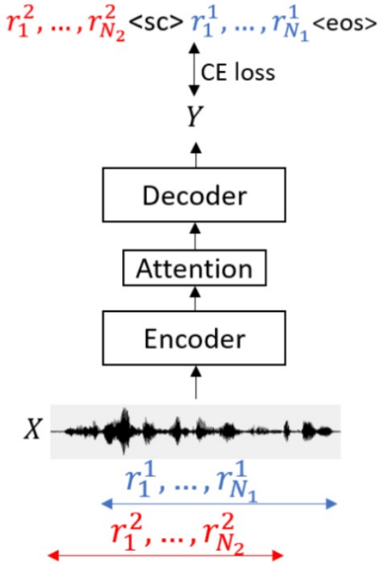
Existing strategy II:

Full end-to-end models

- Usually train from scratch
- Complicated customization



Permutation Invariant Training [3]



Serialized Output Training [4]

[3] Xuankai Chang et al. “End-to-End Multi-speaker Speech Recognition with Transformer,” Interspeech 2020
[4] Naoyuki Naoyuki et al. "Serialized output training for end-to-end overlapped speech recognition," Interspeech 2020

Outline

1. Background
2. **Objective**
3. Proposed Approach
4. Experiments
5. Conclusion

Image source: https://ridermagazine.com/2011/05/13/a-short-history-of-sidecars/

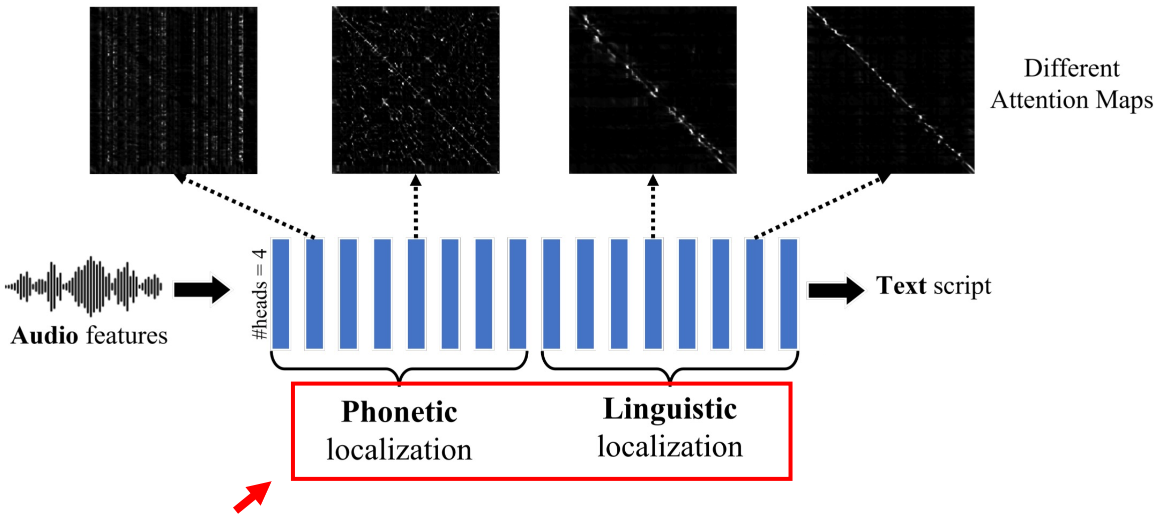
2. Objective

To develop an approach to adapt well-trained common ASR models for multi-talker scenes.

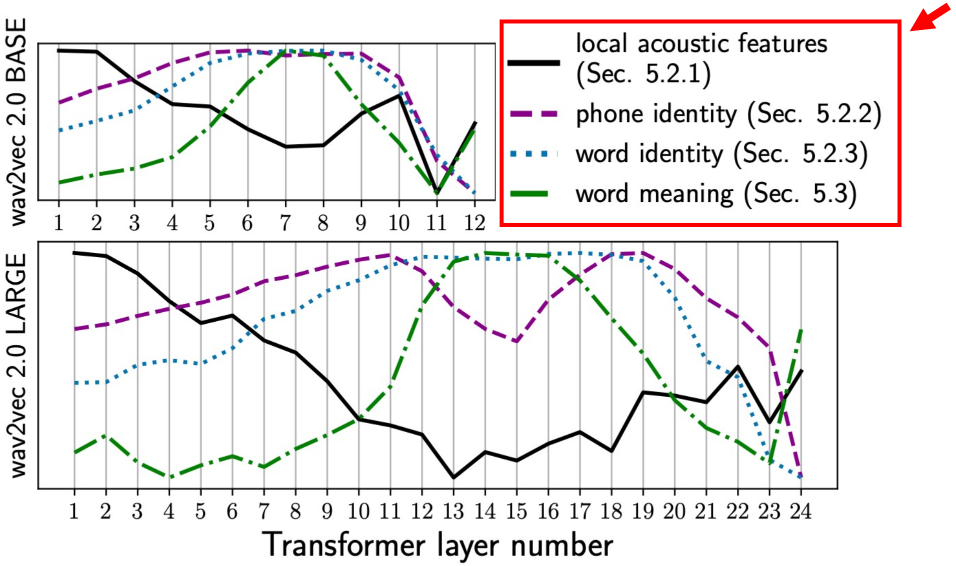
The approach should be low-cost and loose-coupling.

- Low-cost: leverage well-trained models; need only slight training effort
- Loose-coupling: plug-and-play, without distorting original ASR model

2. Objective – Two Inspirations (1/2)



[5]



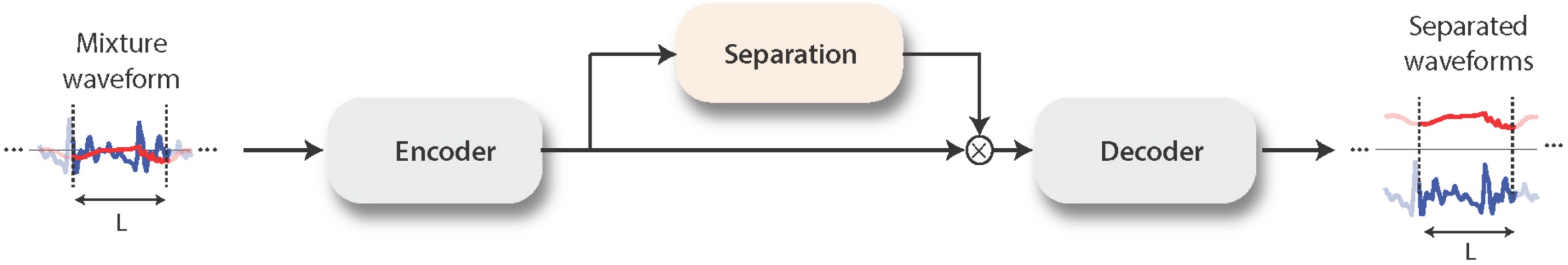
[6]

➢ **Inspired by recent Layer-wise analyses of ASR models**

- Different levels of information are captured with different encoder layers.

[5] Shim, Kyuhong, Jungwook Choi, and Wonyong Sung. "Understanding the role of self attention for efficient speech recognition." ICLR 2022.
[6] Pasad, Ankita, Ju-Chieh Chou, and Karen Livescu. "Layer-wise analysis of a self-supervised speech representation model." IEEE ASRU 2021.

2. Objective – Two Inspirations (2/2)



- **Inspired by methodologies in speech separation**
  - Speech separation usually only involves *low-semantic-level operations*.

[7] Luo, Yi, and Nima Mesgarani. "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation." IEEE/ACM TASLP, 2019.

2. Objective

A potential solution to the objective:

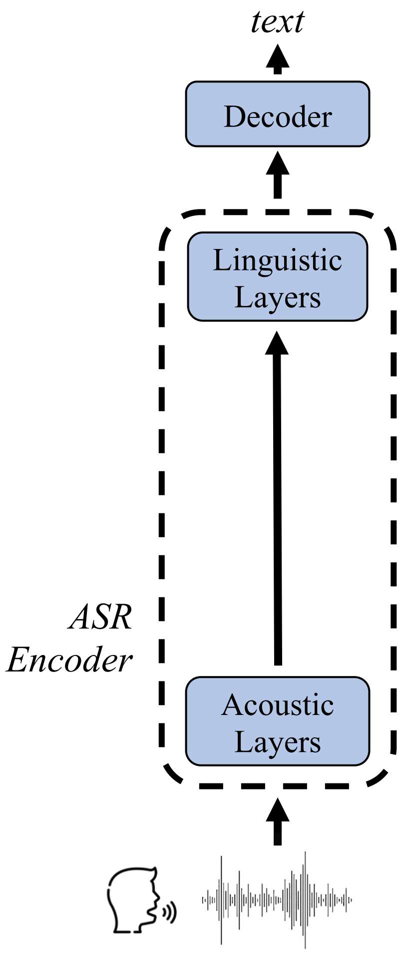
Separate the speech embeddings for different speakers from a lower layer of a well-trained ASR model.

Outline

1. Background
2. Objective
3. **Proposed Approach**
4. Experiments
5. Conclusion

Image source: https://ridermagazine.com/2011/05/13/a-short-history-of-sidecars/

3. Proposed Approach – Multi-talker ASR system with Sidecar

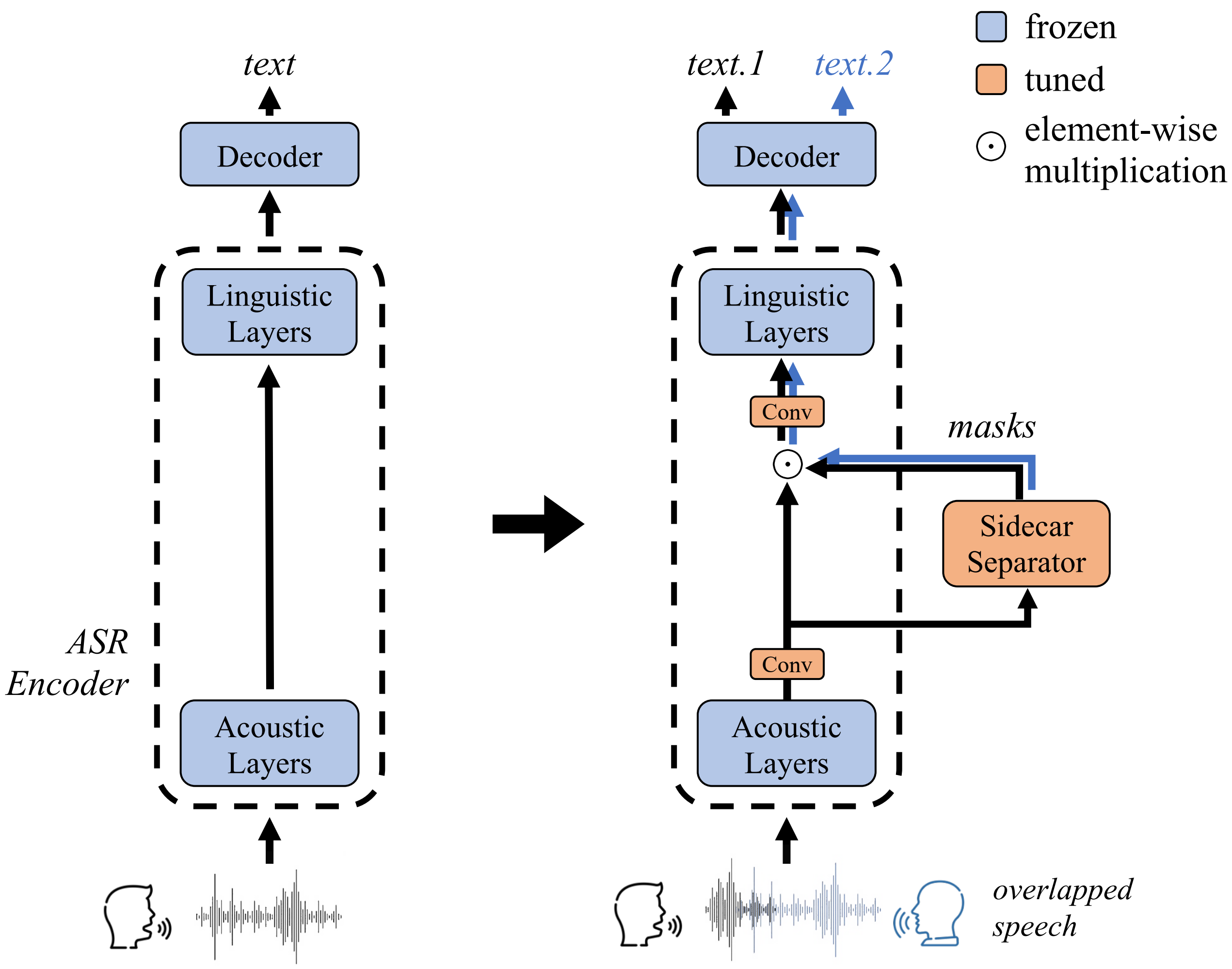


Single-Talker ASR sys.
params: 94.4M

frozen
tuned
⊙ element-wise multiplication

Leverage a well-trained ASR model, whose parameter is frozen.

3. Proposed Approach – Multi-talker ASR system with Sidecar



Single-Talker ASR sys.
params: 94.4M

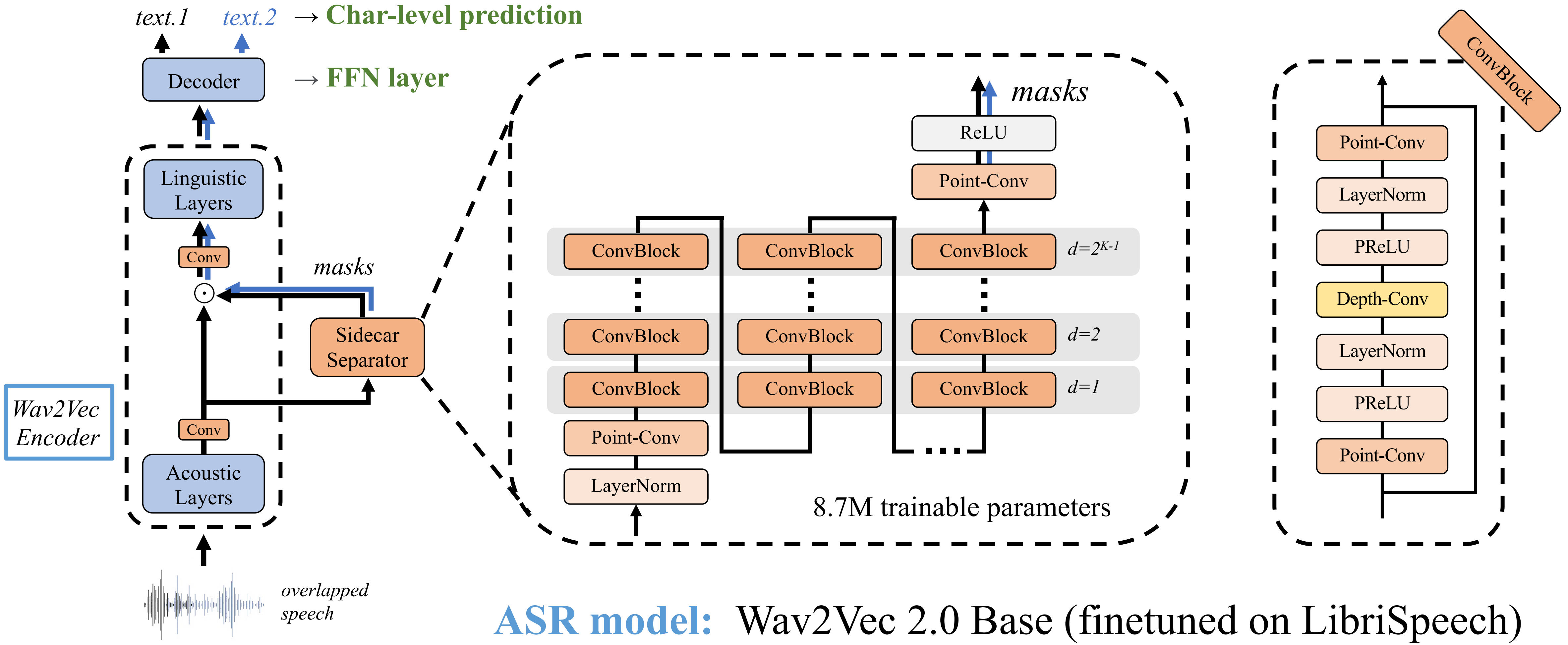
(Sidecar) Multi-talker ASR sys.
params: 103.1M (8.7M trainable)

Leverage a well-trained ASR model, whose parameter is frozen.

Use a "*Sidecar*" to separate speech embeddings. The Sidecar is tunable with ASR loss.

<u>Low-cost</u> and <u>Loose-coupling</u>.

3. Proposed Approach – Detailed implementation

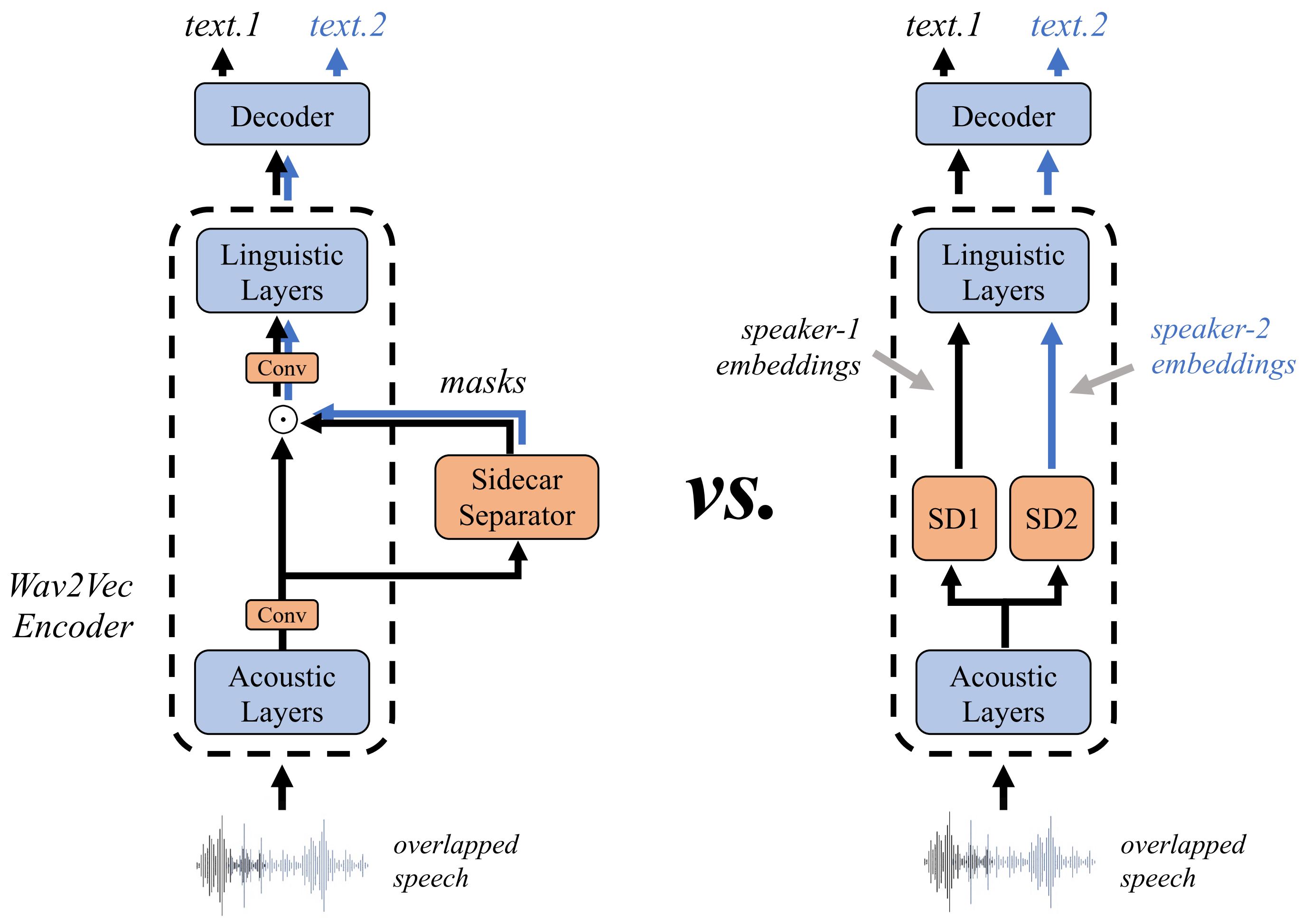


ASR model: Wav2Vec 2.0 Base (finetuned on LibriSpeech)

Sidecar Separator: with a Conv-TasNet-like architecture

Objective Function: CTC loss

3. Proposed Approach– A baseline system for control



(Sidecar) Multi-talker ASR sys.
params: 103.1M (8.7M trainable)

(Baseline) Multi-speaker ASR sys.
#params: 101.5M (14.2M trainable)

To investigate the improvement provided by Sidecar, we also designed a baseline system.

Baseline system:

- Also leverages a well-trained ASR model
- Directly predicts speaker-dependent speech embeddings.

SD: two duplicated layers of the ASR encoder

Outline

1. Background
2. Objective
3. Proposed Approach
4. **Experiments**
5. Conclusion

Image source: https://ridermagazine.com/2011/05/13/a-short-history-of-sidecars/

4. Experiments – LibriMix 2-speaker dataset

LibriMix Dataset: The shorter speech is fully overlapped with the longer one

Systems	**Dev**	**Test**
PIT-Transformer	26.58	26.55
Conditional Conformer	24.50	24.90
ConvTasNet+Transformer	21.00	21.90
DPRNN-TasNet+Transformer	15.30	14.50
Baseline (proposed)	11.60	12.27
W2V-Sidecar (proposed)	**9.76**	**10.36**
W2V-Sidecar (finetune the whole model)	**7.68**	**8.12**

Achieved new state-of-the-art results

4. Experiments – LibriSpeech2Mix 2-speaker dataset

LibriSpeechMix Dataset: The two speech are partially overlapped

Systems	**Dev**	**Test**
PIT-BiLSTM	-	11.1
SOT-BiLSTM	-	11.2
SURT	-	7.2
SOT-transformer	-	5.3
Baseline (proposed)	9.50	9.41
W2V-Sidecar (proposed)	7.76	7.56
W2V-Sidecar (finetune the whole model)	6.01	5.69

Achieved competitive results with far less training effort †

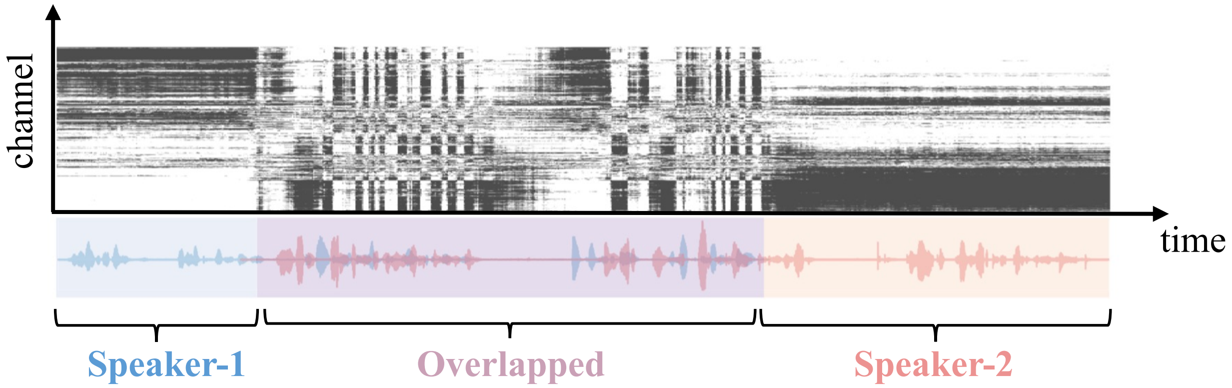
† We trained our model with 8 GPUs for 100k iterations, compared to SOT-transformer's 32 GPUs for 480k.

4. Experiments – Ablation Study

- The Location (in between two encoder layers) of the Sidecar
  - Location 2 (between layers 2 and 3) gave the best performance
    - Intermediate location between lower-layer acoustics and upper-layer linguistics

	Locations							
LibriMix	0	1	2	3	4	6	9	12
Dev	12.18	11.22	**9.76**	12.06	16.14	30.03	56.38	61.78
Test	13.01	11.87	**10.36**	12.65	16.88	30.32	57.11	62.72

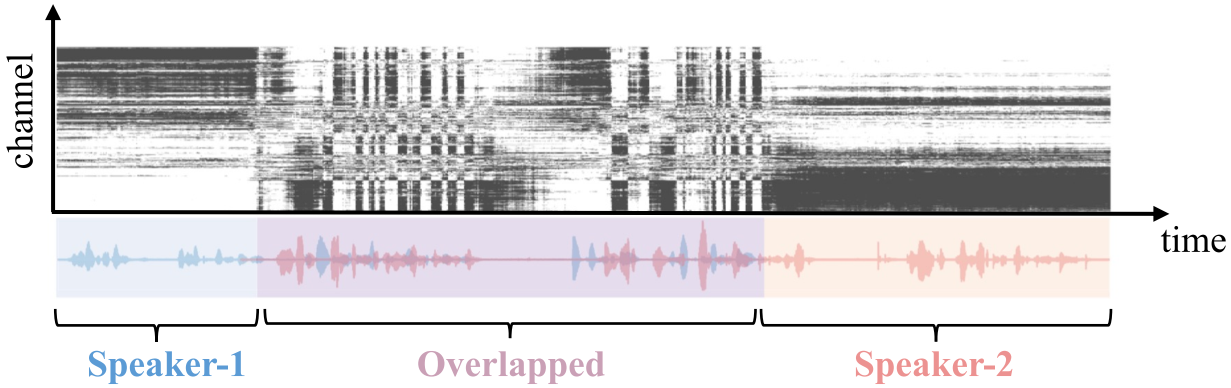
4. Experiments – Visualizations on Sidecar Predicted Masks



Steps of visualizing the masks:

1. Softmax
2. Normalize
3. Cluster

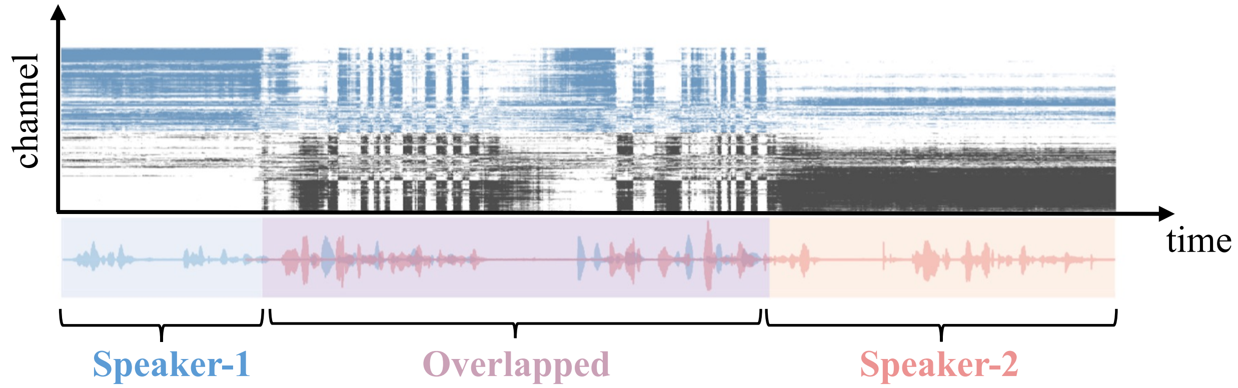
4. Experiments – Visualizations on Sidecar Predicted Masks



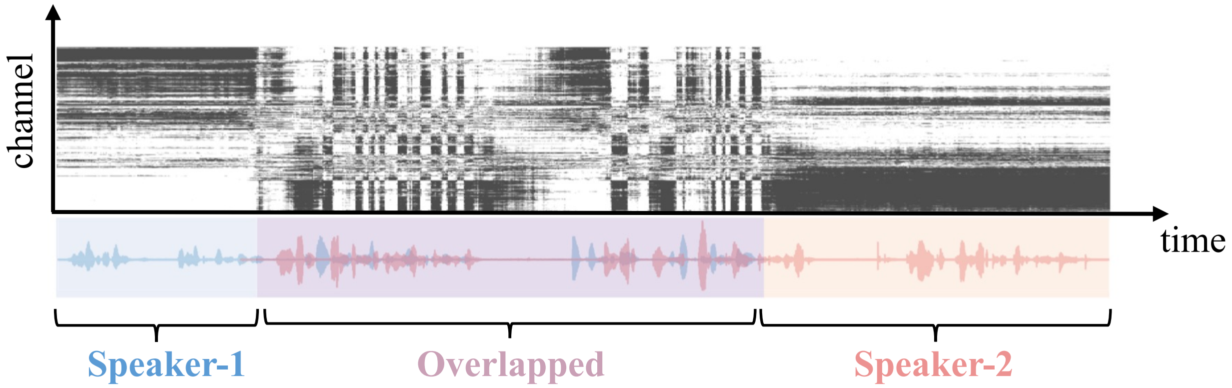
Channel dimension: Sidecar encodes speaker information with different channels

4. Experiments – Visualizations on Sidecar Predicted Masks

Channel dimension: Sidecar encodes speaker information with different channels

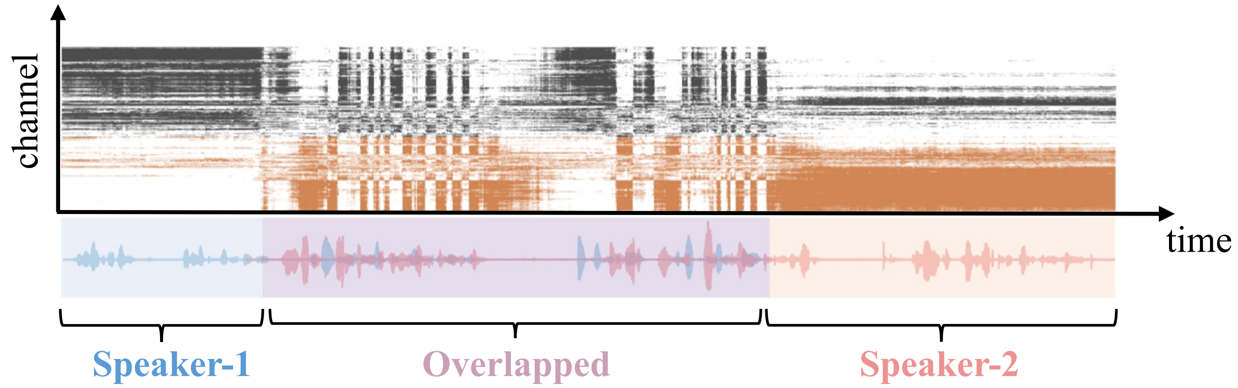


4. Experiments – Visualizations on Sidecar Predicted Masks



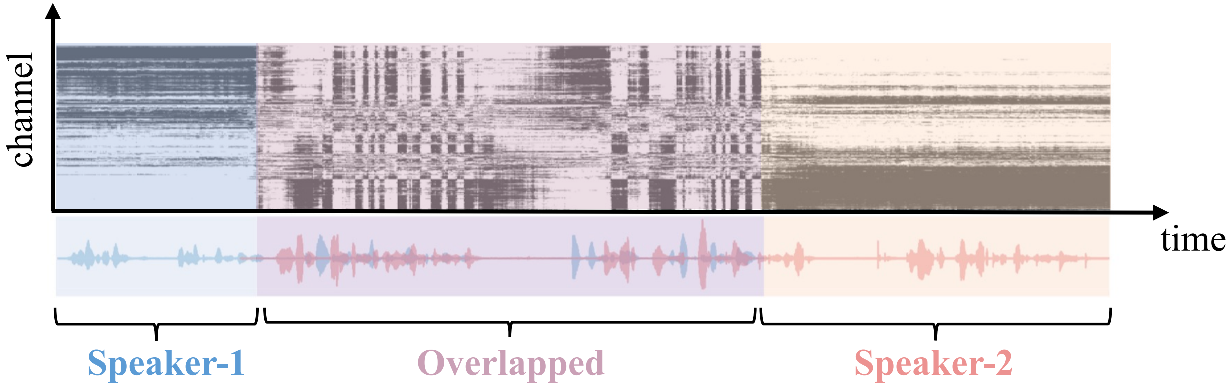
Channel dimension: Sidecar encodes speaker information with different channels

4. Experiments – Visualizations on Sidecar Predicted Masks



Channel dimension: Sidecar encodes speaker information with different channels

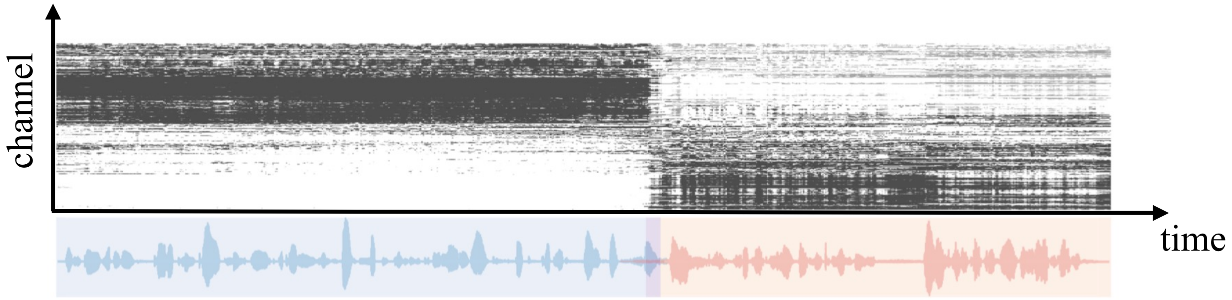
4. Experiments – Visualizations on Sidecar Predicted Masks



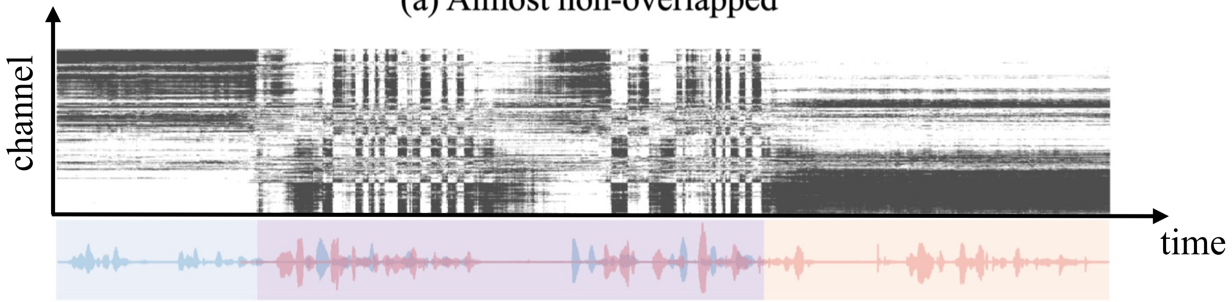
Channel dimension: Sidecar encodes speaker information with different channels

Time dimension: Clear boundary for different part of the utterances

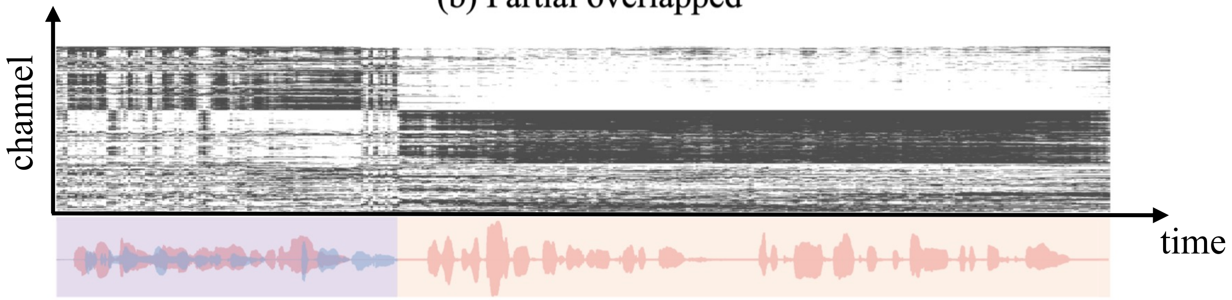
4. Experiments – Visualizations on Sidecar Predicted Masks



(a) Almost non-overlapped



(b) Partial overlapped



(c) Shorter speech is fully overlapped

Channel dimension: Sidecar encodes speaker information with different channels

Time dimension: Clear boundary for different part of the utterances

Speaker diarization?

Outline

1. Background
2. Objective
3. Proposed Approach
4. Experiments
5. **Conclusion**

Image source: https://ridermagazine.com/2011/05/13/a-short-history-of-sidecars/

5. Conclusion

As a multi-talker ASR strategy, Sidecar achieved good performance. It is:

- **Low-cost:** Efficient training , without complicated customization.
- **Loose-coupling:** plug-and-play, without distorting original model’s parameters.

5. Conclusion

As a multi-talker ASR strategy, Sidecar achieved good performance. It is:

- **Low-cost:** Efficient training , without complicated customization.
- **Loose-coupling:** plug-and-play, without distorting original model's parameters.

Further Work:

- Works on 3-spk LibriSpeechMix and LibriMix
- Still works on 1-spk LibriSpeech even trained with multi-speaker

Thank you!

Image source: https://ridermagazine.com/2011/05/13/a-short-history-of-sidecars/