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

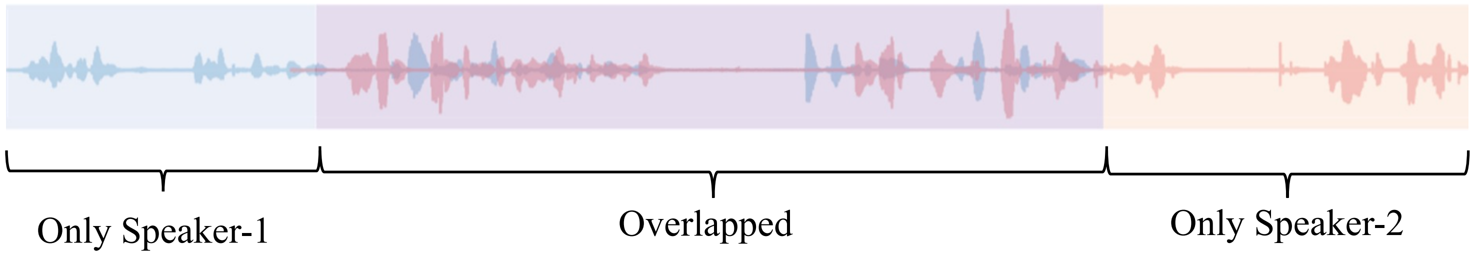
Image source: https://ridermagazine.com/2011/05/13/a-short-history-of-sidecars/

1. **Background**
2. Proposed Method
3. Experiments
4. Conclusion

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

Background - Definition of the task

An example of multi-talker overlapped speech:



Multi-Talker Speech Recognition:

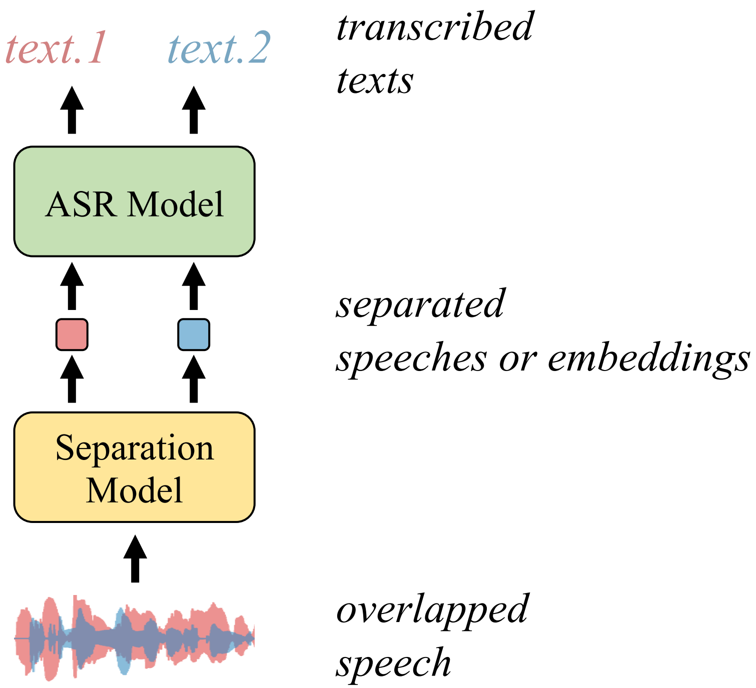
To transcribe texts from multi-talker overlapped speech

Background - Motivations

1. Multi-talker (also known as multi-speaker) speech recognition, where overlapping may exist, remains a challenge.
2. Existing multi-talker ASR strategies have their drawbacks:

Background - Motivations

1. Multi-talker (also known as multi-speaker) speech recognition, where overlapping may exist, remains a challenge.
2. Existing multi-talker ASR strategies have their drawbacks:

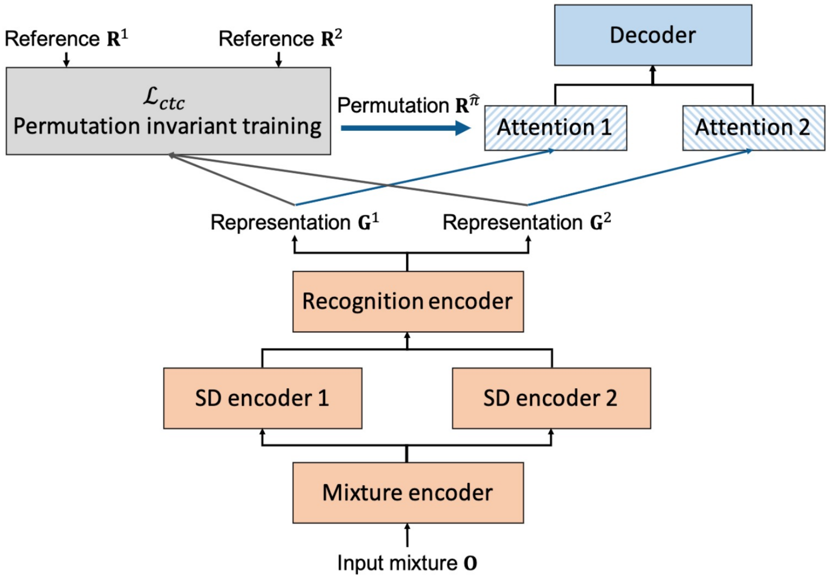


Existing strategy I:
The cascade architectures of a separation model and an ASR model.

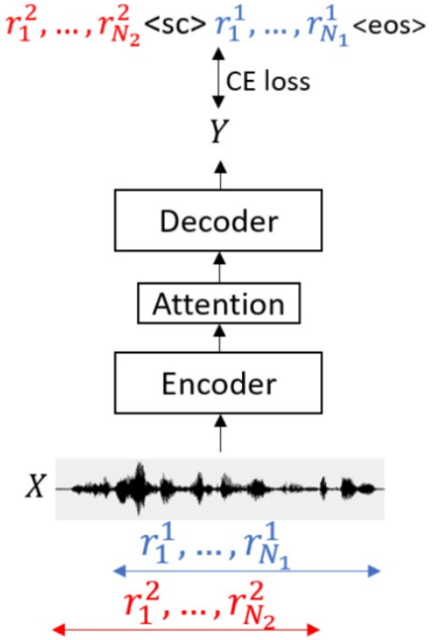
- However, the cascaded modules did not share a uniform training objective, and need further joint fine-tuning.
- The fine-tuned modules cannot work back well in their original domains anymore.

Background - Motivations

1. Multi-talker (also known as multi-speaker) speech recognition, where overlapping may exist, remains a challenge.
2. Existing multi-talker ASR strategies have their drawbacks:



Permutation Invariant Training [1]



Serialized Output Training [2]

Existing strategy II:
Full end-to-end models.

- They usually need training from scratch, and can not take full advantage of the off-the-shelf common ASR systems.
- Some methods need complicated customization.

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

Background - Motivations

2. Existing multi-talker ASR strategies have their drawbacks:

Existing strategy I:
The cascade architectures of a separation model and an ASR model.

- However, the cascaded modules don't share the uniform training objective, and need further jointly fine-tuning.

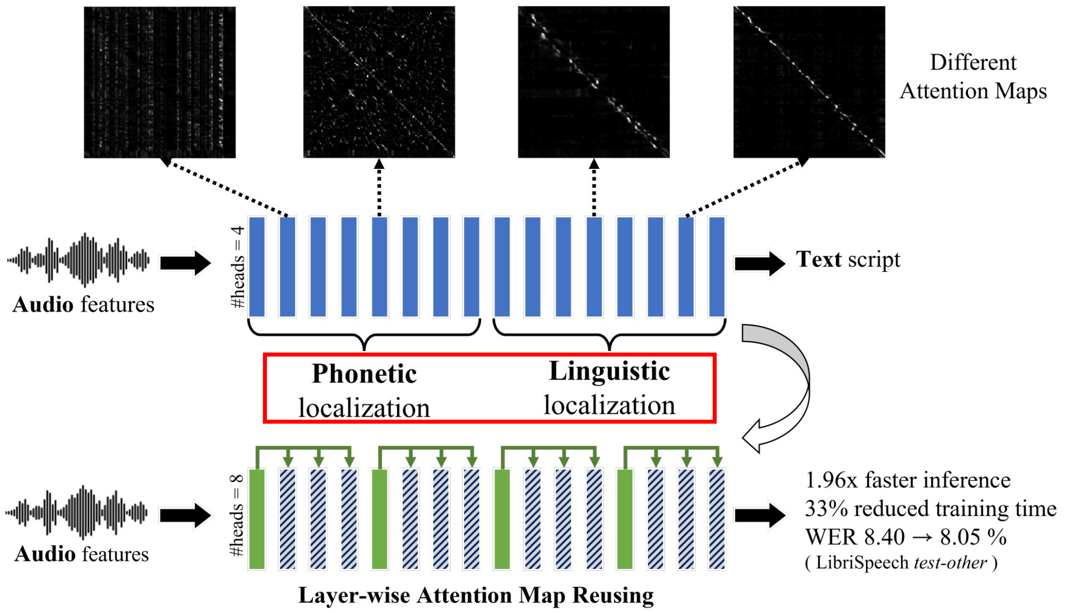
- The fine-tuned modules cannot work well in their original domains anymore.

Existing strategy II:
Fully end-to-end model.

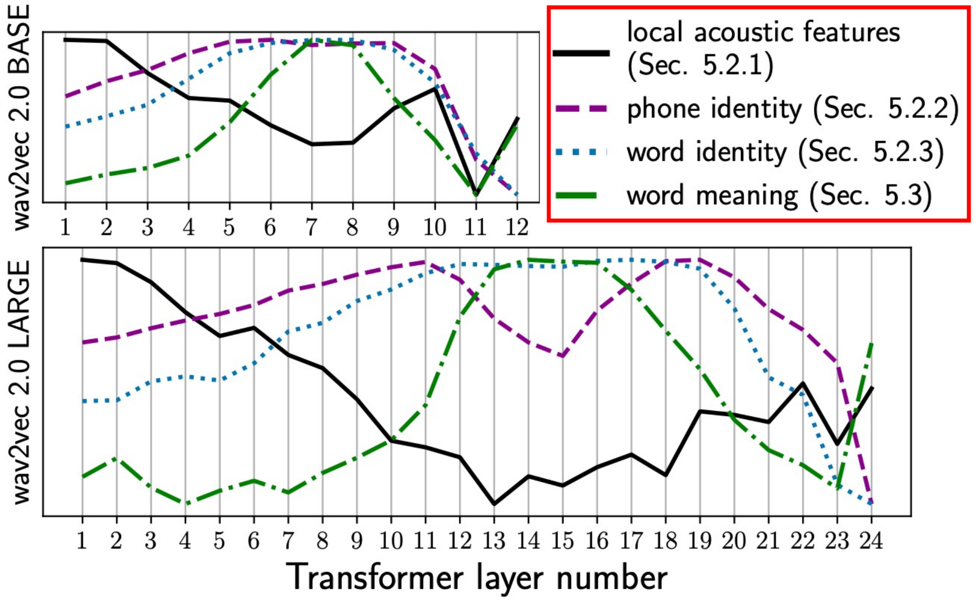
- However, existing methods do not take full advantage of the readily available advancements made for single-speaker ASR.

These drawbacks motivating us to find a low-cost and loose-coupling approach to adapt well-trained single-talker ASR models for multi-talker scenes without distorting the original model's parameters.

Background - Two Inspirations (1/2)



[3]



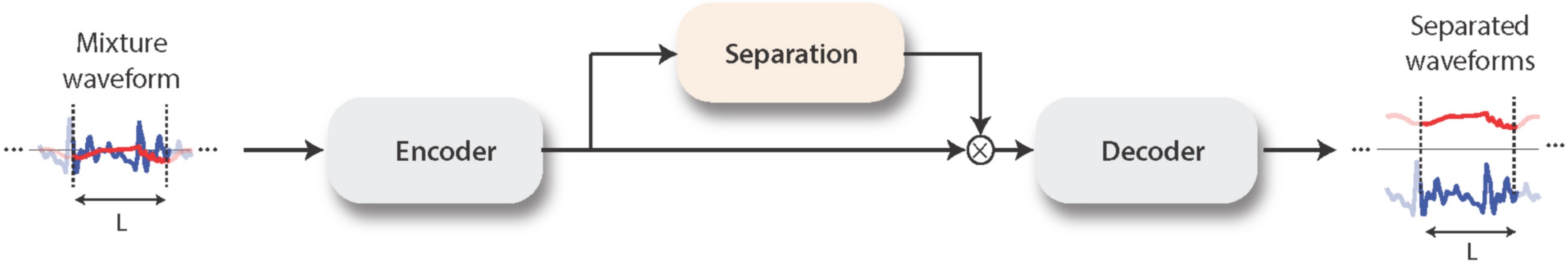
[4]

1. Inspired by recent analyses of ASR models

- Layer-wise analyses of well-trained ASR model indicates that different levels of information are captured with different encoder layers.
- The lower the more acoustic-related (low semantic), and the higher the more linguistic-related (high semantic).

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

Background - Two Inspirations (2/2)



2. **Inspired by methodologies in speech separation**

- Speech separation methods, such as Conv-TasNet, predicts masks for separating mixed speech embeddings. They usually only involves *low-semantic-level operations*.

Drawing on these two inspirations, we hypothesize that within a well-trained ASR encoder, a lower acoustic layer exists where the embeddings of different speakers can be separated.

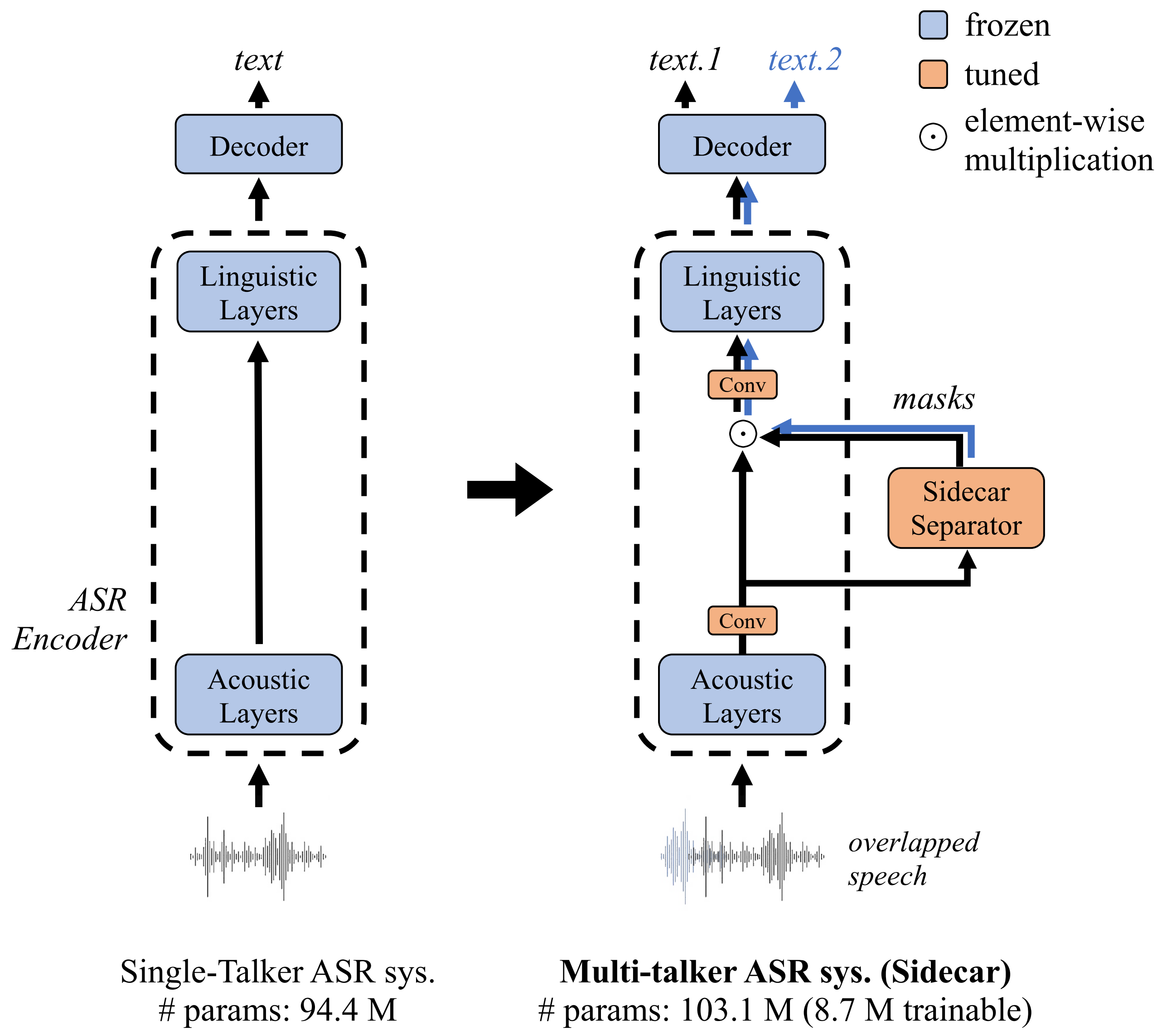
This empowers a common ASR system to handle multi-talker ASR at a low cost in a loose-coupling style.

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

1. Background
2. **Proposed Method**
3. Experiments
4. Conclusion

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

Proposed Method - Multi-talker ASR system with Sidecar



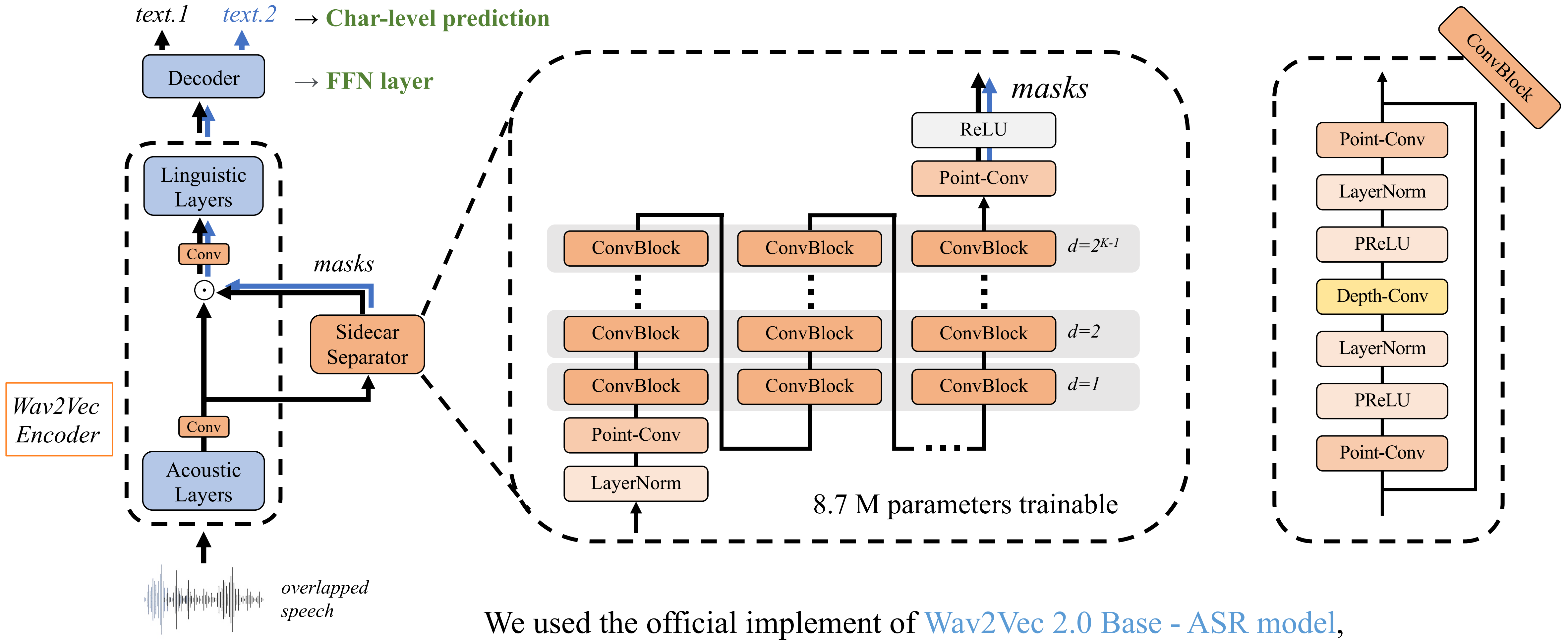
Within a well-trained common ASR system, we separate the mixed speech embedding into multiple speaker-dependent embeddings using a "Sidecar", which located between the encoder's two suitable layers. Two conv layers are plugged in to modulate the embeddings.

The original model is fixed, and only the Sidecar is trainable with ASR loss.

Low-cost: with only slight training efforts
Loose-coupling: without distorting the original model's parameters

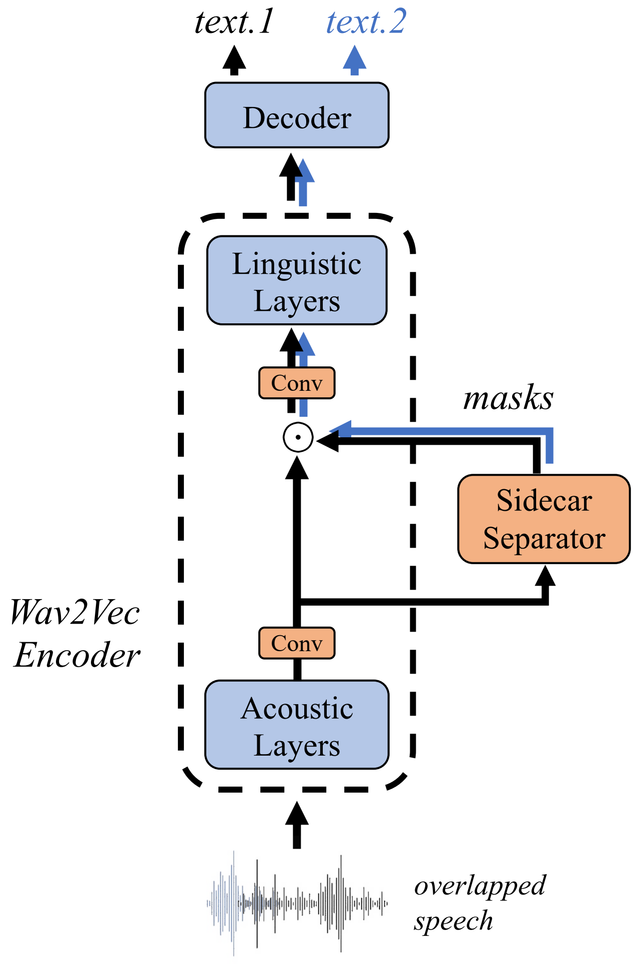
Image source: https://ridermagazine.com/2011/05/13/a-short-history-of-sidecars/

Proposed Method – Detailed implementation



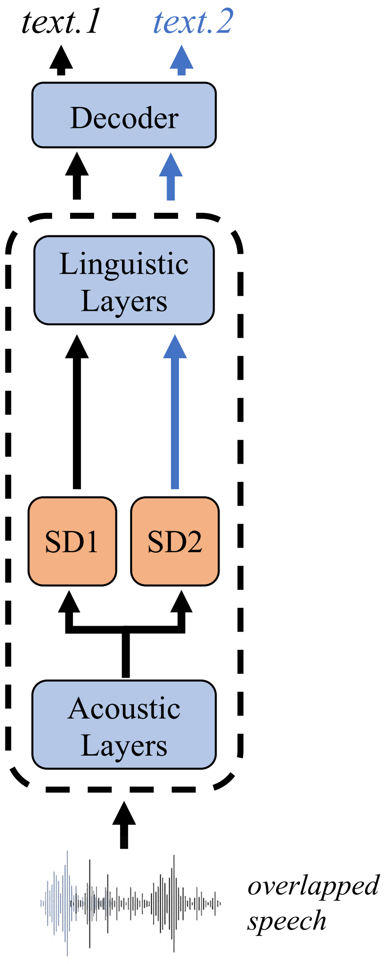
We used the official implement of Wav2Vec 2.0 Base - ASR model,
and used a Conv-TasNet-like architecture as our Sidecar separator,
and used CTC loss for char-level prediction, and an *optional* reconstruction loss.

Proposed Method – A baseline system for control



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

vs.



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

The contribution of a well-trained model is intuitive, while the boost in performance provided by Sidecar can be indistinct.

We also designed a baseline system for control, which also leverages the well-trained ASR model but <u>directly predicts speaker-dependent speech embeddings.</u>

1. Background
2. Proposed Method
3. **Experiments**
4. Conclusion

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

Experiments – Results on 2spk LibriMix and LibriSpeechMix

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

Table 1. Comparison of different systems on *LibriMix*. Evaluated by WER (%). "Transf." refers to "Transformer" and "ft." refers to "fine-tune the whole model".

Systems	Dev	Test
(a) PIT-Transf. [5]	26.58	26.55
(b) Conditional Conformer [30]	24.50	24.90
(c) Con-TasNet + Transf. [5]	21.00	21.90
(d) DPRNN-TasNet + Transf. [5]	15.30	14.50
(e) Baseline (proposed)	11.60	12.27
(f) W2V-Sidecar (proposed)	**9.76**	**10.36**
(g) W2V-Sidecar-ft. (proposed)	**7.68**	**8.12**

Achieved the new state-of-the-art results

LibriSpeechMix Dataset:
The two speeches are partially overlapped

Table 2. Comparisons of different systems on *LibrispeechMix*. Evaluated by WER (%). "-" refers to "not reported" and "ft." refers to "fine-tune the whole model".

Systems	Dev	Test
(a) PIT-BiLSTM [10]	-	11.1
(b) SOT-BiLSTM [10]	-	11.2
(c) SURT [11]	-	$7.2^{\dagger}$
(d) SOT-transf. [31]	-	$5.3^{\dagger}$
(e) Baseline (proposed)	9.50	9.41
(f) W2V-Sidecar (proposed)	7.76	7.56
(g) W2V-Sidecar-ft. (proposed)	6.01	5.69

†With different training data.

Competitive results with far few training effort

(Unlisted) Our subsequent work also demonstrated the effectiveness of this method on 3-spk LibriSpeechMix and LibriMix. We also tested the model, which was trained on the 2-spk data, on the test-clean set of LibriSpeech (1spk), and attained a WER of 3.98%.

Experiments – Ablation studies

Table 3. Ablation study on Sidecar's location, with LibriMix dataset. Results in WER (%).

	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

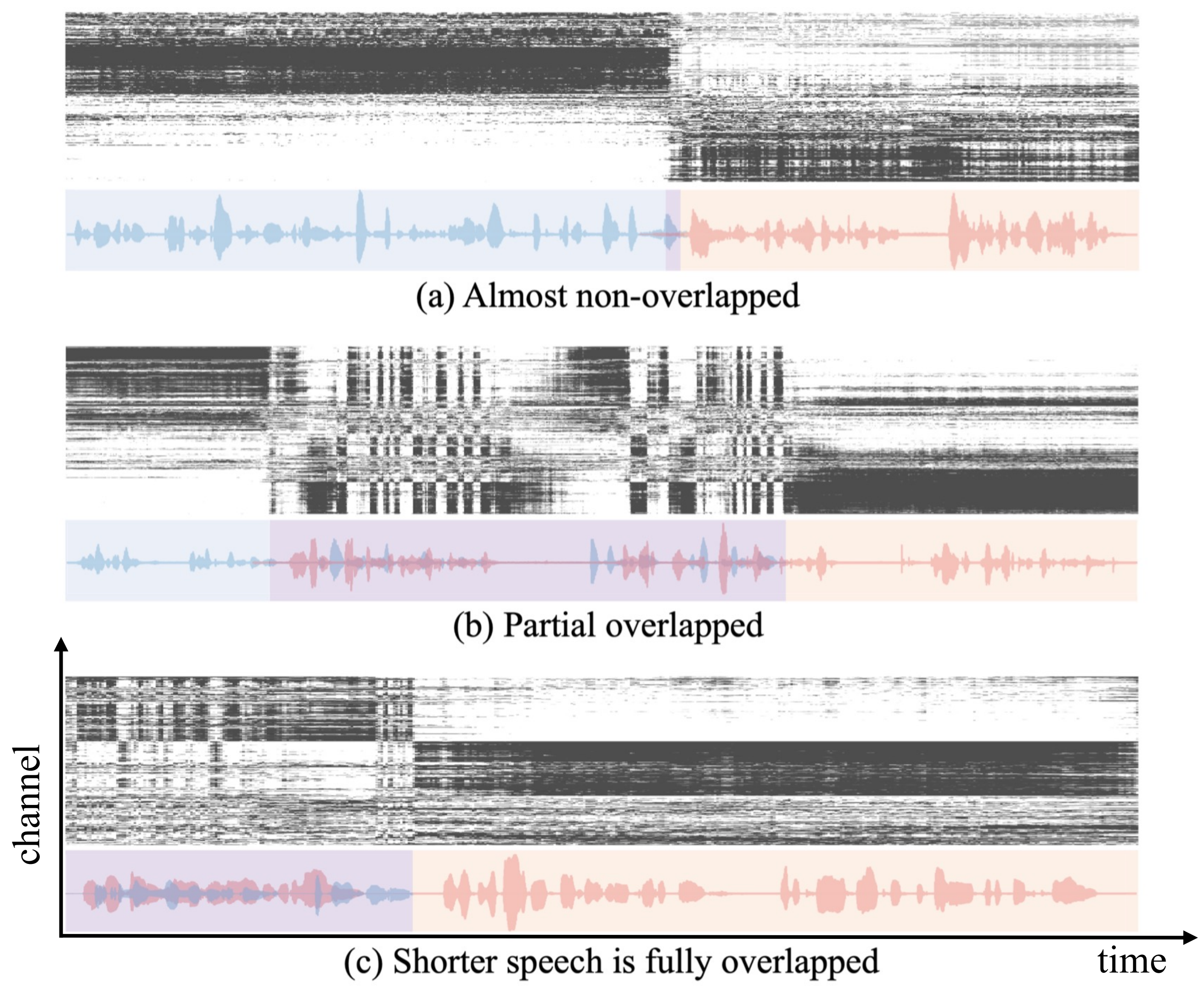
Location 2 is the best.
We believe this is a compromised scale in semantics.

Table 4. Results on WER (%) of with or without reconstruction loss.

	LibriMix		LibriSpeechMix	
	Dev	Test	Dev	Test
W2V-Sidecar	9.76	10.36	7.76	7.56
w/ SISNR	9.69	10.16	7.43	7.20
w/ MSE	9.74	10.32	7.90	7.34

Reconstruction objectives helps slightly, but with high additional training cost.
We did not use reconstruction objectives in our main experiments.

Experiments – Visualization on the Sidecar predicted masks



The visualization indicates that, Sidecar encodes speaker information with different channels and indicates clear distinctions in time domain.

This inspired us to explore the prospects of its application to speech diarization in our future work.

1. Background
2. Proposed Method
3. Experiments
4. **Conclusion**

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

Conclusion

As a multi-talker ASR strategy, Sidecar is

- **Low-cost:**
  Efficient training, without complicated customization.
- **Loose-coupling:**
  the trained Sidecar is plug-and-play without distorting the original model's parameters.
- With **SOTA or competitive performance** with limited training.

Thank you!

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