A SIDECAR SEPARATOR CAN CONVERT A SINGLE-SPEAKER SPEECH RECOGNITION SYSTEM TO A MULTI-SPEAKER ONE

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

The Chinese University of Hong Kong, Hong Kong SAR, China

{lmeng, jwkang, mycui, wangy, wuxx, hmmeng}@se.cuhk.edu.hk

ABSTRACT

Although automatic speech recognition (ASR) can perform well in common non-overlapping environments, sustaining performance in multi-speaker overlapping speech recognition remains challenging. Recent research revealed that ASR model's encoder captures different levels of information with different layers – the lower layers tend to have more acoustic information, and the upper layers more linguistic. This inspires us to develop a *Sidecar* separator to empower a well-trained ASR model for multi-speaker scenarios by separating the mixed speech embedding between two suitable layers. We experimented with a wav2vec 2.0-based ASR model with a Sidecar mounted. By freezing the parameters of the original model and training only the Sidecar (8.7 M, 8.4% of all parameters), the proposed approach outperforms the previous state-of-the-art by a large margin for the 2-speaker mixed LibriMix dataset, reaching a word error rate (WER) of 10.36%; and obtains comparable results (7.56%) for LibriSpeechMix dataset when limited training.

***Index Terms*—** multi-speaker speech recognition, speech separation, end-to-end speech recognition, domain adaptation

1. INTRODUCTION

End-to-end automatic speech recognition (ASR) for common non-overlapping environments has made significant progress recently [1, 2]. However, multi-speaker speech recognition, where overlapping may exist, still remains a challenge [3]. There are two mainstream end-to-end paradigms that aim to tackle the challenge: (i) cascade architectures that jointly fine-tune speech separation and speech recognition modules [4, 5]; and (ii) complete end-to-end models customized deliberately for multi-speaker overlapping speech scenarios [6, 7, 8, 9, 10, 11, 12]. However, the former approach may see performance degradation in modules' original domains, and the latter does not take full advantage of the readily available advancements made for single-speaker ASR. This motivates us to find a low-cost and loose-coupling approach to adapt well-trained single-speaker ASR models for multi-speaker scenes without distorting the original model's parameters.

Recent research investigated the information captured in the layers within the encoders of ASR models. Shim et al. [13] found that the transformer-based encoder extracts acoustic representations in its lower layers, and linguistic representations in the upper layers. Layer-wise analyses of self-supervised speech representation models, such as wav2vec 2.0 [14], proved that they encode representations following an acoustic-linguistic hierarchy from lower to upper layers as well [15, 16], which was further discussed in the context of neuroscience [17, 18]. Similar findings have also been reported for CNN- / RNN-based models [19, 20].

Enlightened by the above findings, we assume that there exists a lower suitable location between the encoder's two layers where the multi-speaker acoustic embedding can be well-separated by drawing on speech separation techniques. In speech separation, research demonstrated that predicting masks for separation is usually superior to directly predicting separated embeddings [21, 22]. As a representative, the TasNet architecture [23] predicts masks in the time domain for mixed speech embeddings and achieves impressive results. Subsequently, Luo et al. [24] proposed the well-recognized Conv-TasNet, which predicts masks using a convolutional neural network, which made a further leap in performance.

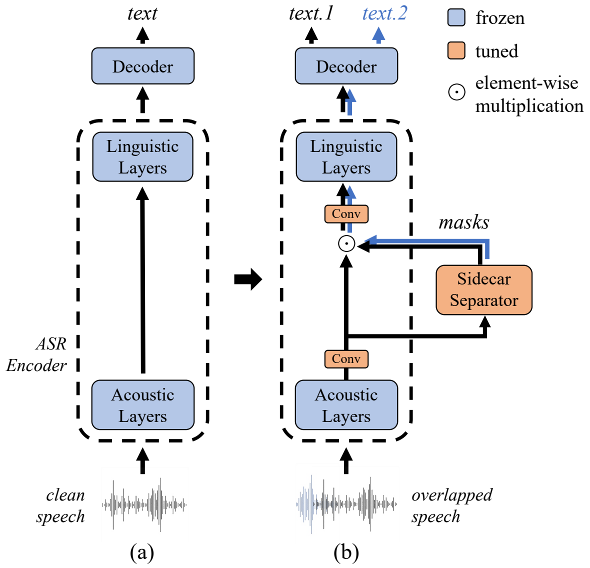


Fig. 1. (a) a well-trained single-speaker ASR system; (b) the proposed strategy with a Sidecar separator mounted, taking 2 speakers as example. On both sides of Sidecar, we add a convolutional layer to coordinate its input-overlapping and output-separated embeddings.

Inspired by the recent analyses of ASR models and methodologies in speech separation, we introduce a promising strategy to adapt off-the-shelf well-trained ASR models to multi-speaker scenarios. As shown in Fig. 1, we mount a *Sidecar* separator between two suitable layers of a well-trained ASR model. On both sides of the Sidecar, there is a simple convolutional layer helping to coordinate the input-overlapping and output-separated embeddings of the Sidecar. The proposed approach has three key advantages:

- The approach is low-cost and loose-coupling for converting a well-trained single-speaker ASR model to a multi-speaker one, without complicated customization on the model structure or on the training scheme.
- The original ASR model is well-trained and fixed, and only Sidecar (8.7 M, 8.4% of all parameters) needs tuning, making the training feasible within limited time and GPU resources.

- Experiments leveraging a wav2vec 2.0-based ASR model mounted with a Sidecar are conducted, achieving a WER of 10.36% on 2-speaker LibriMix dataset and 7.56% on LibriSpeechMix dataset with limited training.

Moreover, by visualizing Sidecar-predicted masks (Fig. 3), we find that among channel dimensions different features encode different speakers' information. And in the time domain, there exist clear distinctions between the periods of speech for different speakers, and the periods of overlapping speech.

2. MULTI-SPEAKER ASR SYSTEM WITH SIDECAR

The proposed methodology consists of three main components — a well-trained single-speaker ASR model with parameters frozen, a Sidecar separator, and the training objective. As shown in Fig. 1, with the cooperation of two convolutional layers, Sidecar is plugged between two layers of ASR encoder and forms a multi-speaker ASR system. No language models or lexicons are used in this work.

With permutation invariant training (PIT) [25], the model is optimized using connectionist temporal classification (CTC) loss [26].

2.1. Well-trained single-speaker ASR model

A typical end-to-end ASR model contains an encoder to synthesize waveform or acoustic features into high-level representations, and a decoder to model the representations into language tokens. It often takes much time and effort to train an ASR model from scratch, let alone in multi-speaker environments. With many off-the-shelf single-speaker models already available, we try to reuse the single-speaker model on multi-speaker overlapping speech recognition.

Wav2vec 2.0 [14] is a well-recognized pre-trained speech representation model based on self-supervised learning (SSL), attracting interest in the field. Many ASR models taking wav2vec 2.0 as the encoder reported state-of-the-art performance [27].

In our experiments, we use a well-trained wav2vec 2.0 base-based ASR model, the same as used in [14], which contains a CNN feature extractor, a Transformer encoder, and a fully-connected layer as the decoder. Specifically, the model takes a waveform as input and extracts acoustic features with a 7-layer CNN feature extractor. After a linear projection, the features are fed into the encoder consisting of 12 layers of Transformer blocks for generating high-level representations. We obey the paradigm in [14] and only use a fully-connected layer as the decoder for letter-level prediction.

We directly use the model parameters released by fairseq [28], denoted as *W2V-CTC* in the following parts.

2.2. Sidecar separator

Inspired by the findings that the ASR encoder captures more acoustic information in its lower layers and more linguistic information in the upper layers [13, 15, 20], we propose using a Sidecar separator to address multi-speaker speech recognition, drawing on methodologies in speech separation.

As shown in Fig. 2, similar to Conv-TasNet [24], the Sidecar is a temporal convolutional network consisting of stacked 1-D dilated convolutional blocks, which allows the Sidecar to model the long-term dependencies of the acoustic embeddings while maintaining a small size. By ablation study, we plug the Sidecar into the most suitable position between two "acoustic" layers of the ASR encoder. Since the original ASR model is frozen, to alleviate the "transplant rejection", we use a 3-kernel-size convolutional layer to filter Sidecar's input-mixed and output-separated embeddings, respectively.

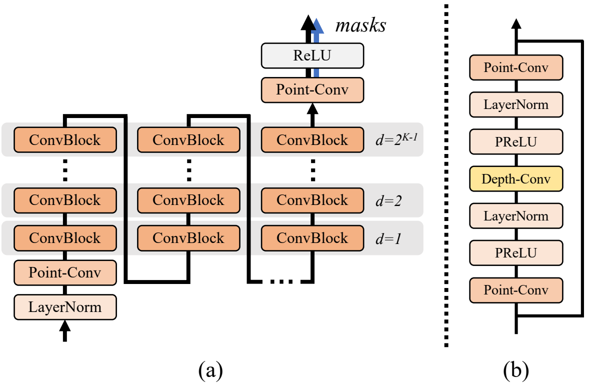


Fig. 2. (a) Sidecar structure; (b) Details in ConvBlock. Referring to Conv-TasNet [24], the Sidecar consists of stacked 1-D dilated convolutional blocks. d represents the dilation rate.

In the forward process, the preceding-layer-generated mixed speech embedding is filtered by a convolutional layer, and fed into the Sidecar to generate the speaker-dependent masks. Then, the filtered mixed speech embedding will be element-wise multiplied by the masks, and further adjusted by another convolutional layer to obtain the separated embeddings. The separated embeddings of different speakers will go in parallel through the rest of the model and be transcribed into text. This is technically implemented by concatenating the separated embeddings onto the batch dimension.

2.3. Training objective

We favor the use of permutation invariant training (PIT) with only ASR loss, which is CTC loss in this work. This can already achieve satisfactory performance.

In addition to PIT-CTC loss, we also tried two reconstruction objectives: maximize scale-invariant signal-to-noise ratio (SI-SNR) or minimize mean squared error (MSE). Since the multi-speaker dataset is simulated from single-speaker speeches, reconstruction loss aims to drive the predicted separated embeddings as close as possible to corresponding clean single-speaker embeddings. The clean single-speaker embeddings are generated by the transformer layer before where the Sidecar is plugged in, and the permutation of speakers is determined by PIT-CTC loss.

Although introducing a reconstruction loss can provide a minor performance gain (Table 4), we do not recommend this. Because it requires input not only mixed speech but also clean single-speaker speeches, which significantly increases the computational burden.

3. A BASELINE SYSTEM FOR CONTROL

We attribute the effectiveness of this work to two aspects: the knowledge of the well-trained single-speaker ASR model, and the Sidecar which can efficiently adapt the former to multi-speaker scenarios by predicting speaker-dependent masks rather than their embeddings. The contribution of a well-trained model is intuitive, while the boost in performance provided by Sidecar can be indistinct.

Considering this, we designed a baseline system, which also leverages the same ASR model as our proposed approach, but directly predicts speaker-dependent speech embeddings like [8]. Specifically, in the same position as Sidecar is in our proposed approach, the baseline model duplicates the preceding encoder layer to predict speaker-dependent embeddings. Except for the two duplicated layers, other parameters are frozen.

Note that, unlike our proposed approach, this Baseline does not maintain the property of keeping the original model parameters unchanged, because it fine-tunes the layers of the original model.

4. EXPERIMENTAL SETUP

4.1. Datasets

The experiments are performed on two benchmark datasets: *LibriMix* [29] and *LibriSpeechMix* [10]. Both of them are simulated from *LibriSpeech* dataset but with different protocols.

LibriMix. It is simulated with the mixtures of two or three speakers, in a clean or noisy environment. We focus on its 2-speaker-clean subset *Libri2Mix-clean*. *Libri2Mix-clean*'s training, development, and test set contain 270 hours, 11 hours, and 11 hours of 2-speaker's mixed speeches, respectively. The mixtures are made in a left-aligned style. Therefore, the shorter source speech will be fully overlapped by the longer one from the beginning.

LibriSpeechMix. It only has standard dev and test sets. We focus on the 2-speaker "*dev-clean-2mix*" and "*test-clean-2mix*" for validation and test. The 2-speaker training set is homemade from the 960-hour LibriSpeech training dataset (LS-960) using the same protocol as [10]. LibriSpeechMix randomly samples a delay time for a second utterance so the mixture is partially overlapping.

Compared with LibriSpeechMix, LibriMix has larger overlap ratios, which greatly challenges the model in separating overlaps.

4.2. Model settings

Well-trained single-speaker ASR model.

In accordance with the paradigm in [14], the used W2V-CTC model contains a CNN feature extractor, a Transformer encoder, and a fully-connected layer as the decoder. We directly reuse the parameters released by Fairseq[1] [28], which was first pre-trained on unlabeled LS-960 with contrastive loss and diversity loss, then fine-tuned on labeled LS-960 with CTC loss. It reached a WER of 3.4% on LibriSpeech test-clean dataset, and 8.9% on test-other dataset.

Sidecar separator. Referring to [24], in our Sidecar separator, K convolutional blocks with dilation rates 1, 2, 4, ..., 2^{K-1} are repeated R times. We take $K = 8, R = 3$, which performs better. We discard skip-connection paths of convolutional blocks, and change the final sigmoid activation to ReLU to fit our task. The Sidecar uses 128 bottleneck-channels, and 768 input- / output- channels. Ablation experiments (Section 5.4) have been conducted to explore the most suitable Sidecar location. As a result, we plug the Sidecar right after the second transformer layer and before the third, which gave the best performance. With W2V-CTC frozen, it only has 8.7 M (8.4% of all parameters, about half of the Baseline) for tuning.

Training settings. We optimize the proposed model and Baseline using a 2e-4 learning rate with a three-stage schedule and Adam optimizer, for at most 100 k updates. It takes about 7 hours for models' convergence with 8 NVIDIA V100 GPUs, thanks to Sidecar's small size and the ejection start provided by the well-trained model.

In the following, we denote the proposed model as *W2V-Sidecar*.

5. RESULTS AND DISCUSSION

5.1. Results On LibriMix dataset

We compared different systems on the two benchmark datasets. The corresponding results are shown in Table 1 and Table 2.

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

System	Dev	Test
(a) PIT-Transf. [5]	26.58	26.55
(b) Conditional Conformer [30]	24.50	24.90
(c) Conv-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**

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

System	Dev	Test
(a) PIT-BiLSTM [10]	-	11.10
(b) SOT-BiLSTM [10]	-	11.20
(c) SURT-non-streaming [11]	-	7.20†
(d) SOT-transf. [31]	-	5.30†
(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 heavier training data.

For 2-speaker LibriMix (Table 1), the designed Baseline (e) for control already outperforms previous methods by a large margin. We attribute this improvement to the knowledge of the well-trained model. Then, by comparing the proposed W2V-Sidecar (f) with Baseline (e), we find the introduction of Sidecar further boosts the WER with even less trainable parameters (8.7 M, 8.4% of all parameters, about half of Baseline). This confirms that predicting masks is more effective than directly separating embeddings as discussed in [22]. Besides, the Sidecar serves in a plug-and-play style without distorting the original parameters. This low-coupling property allows the model to be more flexible for multiple scenarios.

Moreover, as an option, we also train a W2V-Sidecar-ft (g), which fine-tunes all model parameters. The training settings are the same as (e). Since the model is initialized with well-trained parameters, the W2V-Sidecar-ft's convergence is also fast. Not surprisingly, the model achieves even more impressive results.

5.2. Results On LibriSpeechMix dataset

For the comparison on the LibriSpeechMix dataset, we only focus on those non-streaming models with no additional auxiliary such as speaker labels or additional training datasets.

As shown in Table 2, our Baseline (e) achieves better performance than (a) and (b) even with fewer training efforts. As a similar trend, the Sidecar (f) brings a further performance boost. And as an option, the W2V-Sidecar-ft (g) reaches a better result at the cost of losing the loose-coupling property. We also list the results of Systems (c) and (d), which with different setups, for a comprehensive comparison. They gain further improvements from their significantly heavier training efforts.

Compared with (a)-(d), the proposed systems (e)-(f) are trained efficiently with only 7 hours using 8 GPUs. Moreover, systems (a), (b), and (d) have larger model sizes than the proposed W2V-Sidecar (94.4 M frozen + 8.7 M trainable).

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

5.3. Visualization of Sidecar predicted masks

To better understand what Sidecar has learned, we investigate its generated masks with visualization. For a better view, we perform element-wise softmax on the two masks derived from a mixed speech embedding to highlight pairwise differences. This process produces two essentially identical matrices, and we just take one. Then we normalize each channel (or feature) using its mean and standard deviation along the time steps to avoid swamping those channels which with minor differences between the two masks. Afterward, we reorder the channel dimension according to hierarchical clustering based on the pairwise distances of the channels.

We randomly take three typical cases as examples: an almost non-overlapped case, a partially overlapped case, and a case in which the shorter speech is fully overlapped. We find interesting clues from their masks' visualization (Fig. 3). The masks show strong temporal correlations with the input speech, and distinctly tell the mixture boundaries. This indicates that different feature sub-spaces (or channel groups) capture different speakers' information, so the speaker boundaries emerge when close channels are clustered.

5.4. Ablation studies

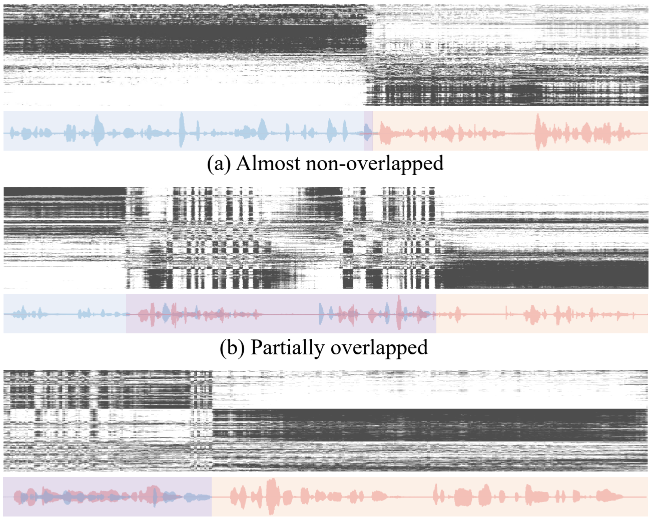
Sidecar location. We explored the optimal location for the Sidecar. We mark the locations by the transformer blocks index. E.g., location **0** is right before the first block; location **1** is between the first and second blocks, etc. The results are summarized in Table 3. Among the locations, the result of location **2** exceeds the previous layers, and deeper locations show a dramatic decrease in results.

This trend is mutually supportive of existing layer-wise analysis research [13, 15], and aligns with our previous hypothesis: the separation is better performed on acoustic-related representations, which contain sufficient low-semantics information to catch phonetic-level differences. We argue that location **0** is too close to the raw input, where meaningful phonetic-level representations have yet to be well-synthesized. Meanwhile, speaker information matters, as the discussion about Fig. 3. However, if the ASR encoder goes deeper, the speaker information will be eliminated. As a result, location **2** is a compromised scale in semantics.

Reconstruction loss. As reconstruction loss plays a dominant role in speech separation tasks, a natural question is whether introducing reconstruction loss can help our task. In this part, we explored adding MSE or SI-SNR loss into W2V-Sidecar. In Table 4, our experiments show that introducing reconstruction loss can make slight improvements. However, the computational burden is significantly increased because it requires not only mixed speech input but also clean single-speaker recordings. We argue that adding a constraint on low-level embeddings for such a high-semantic task may not be very helpful because the mapping can be an ill-posed problem.

5.5. Limitations and future work

This work has several limitations. First, although we used PIT as our training scheme, our strategy also naturally fits serialized output training (SOT), which usually needs to train from scratch. We are interested in whether Sidecar can accelerate SOT's training with a well-trained ASR model. Second, according to Fig. 3, Sidecar explicitly encodes speaker information. We are excited about the prospects of its application to speech diarization, especially combined with SOT. Third, we only implement Sidecar on the ASR task. We also expect its applicability to other downstream tasks when overlapping exists. Finally, to adopt a generally popular speech representation model, we only use wav2vec 2.0-based model. We will try the Sidecar on other SSL or non-SSL models in the future.



(a) Almost non-overlapped

(b) Partially overlapped

(c) Shorter speech is fully overlapped

Fig. 3. Visualization of generated masks and input waveforms. We use different colors to distinguish speakers. Purple represents overlapping. The horizontal of masks is time dimension, and the vertical is channel dimension. Sidecar encodes speaker information with different channels and indicates clear distinctions in time domain.

Table 3. Ablation study on Sidecar location, evaluated by WER (%).

	Location							
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

Table 4. Ablation study on using reconstruction loss, by WER (%).

	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

6. CONCLUSION

Inspired by the findings that ASR encoder captures more acoustic representations in its lower layers and more linguistic in the upper layers, we propose plugging a Sidecar separator into a well-trained single-speaker ASR model and converting it to a multi-speaker one. The original ASR model is frozen, and only 8.4% of all parameters need tuning. With efficient training, the proposed method outperforms previous state-of-the-art by a large margin on the 2-speaker mixed LibriMix dataset, reaching a WER of 10.36% dataset; and comparable results (7.56%) on the LibriSpeechMix dataset.

Visualizations of Sidecar-predicted masks indicate that in channel dimension, different features encode different speakers' information. And in the time domain, there exist significant distinctions between different speaker speech periods and overlapping periods.

7. ACKNOWLEDGEMENT

This research is partially supported by the HKSARG Research Grants Council's Theme-based Research Grant Scheme (Project No. T45-407/19N) and by the CUHK Stanley Ho Big Data Decision Analytics Research Centre.

8. REFERENCES

[1] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE ICASSP*, 2016.

[2] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020.

[3] Zhongxin Bai and Xiao-Lei Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, 2021.

[4] Shane Settle, Jonathan Le Roux, Takaaki Hori, Shinji Watanabe, and John R Hershey, "End-to-end multi-speaker speech recognition," in *IEEE ICASSP*, 2018.

[5] Song Li, Beibei Ouyang, Fuchuan Tong, Dexin Liao, Lin Li, and Qingyang Hong, "Real-time end-to-end monaural multi-speaker speech recognition," in *Interspeech*, 2021.

[6] Hiroshi Seki, Takaaki Hori, Shinji Watanabe, Jonathan Le Roux, and John R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *ACL*, 2018.

[7] Wangyou Zhang, Xuankai Chang, Yanmin Qian, and Shinji Watanabe, "Improving end-to-end single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[8] Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *IEEE ICASSP*, 2020.

[9] Anshuman Tripathi, Han Lu, and Hasim Sak, "End-to-end multi-talker overlapping speech recognition," in *IEEE ICASSP*, 2020.

[10] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," in *Interspeech*, 2020.

[11] Liang Lu, Naoyuki Kanda, Jinyu Li, and Yifan Gong, "Streaming multi-talker speech recognition with joint speaker identification," in *Interspeech*, 2021.

[12] Naoyuki Kanda, Jian Wu, Yu Wu, Xiong Xiao, Zhong Meng, Xiaofei Wang, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Takuya Yoshioka, "Streaming multi-talker ASR with token-level serialized output training," in *Interspeech*, 2022.

[13] Kyuhong Shim, Jungwook Choi, and Wonyong Sung, "Understanding the role of self attention for efficient speech recognition," in *Proceedings of International Conference on Machine learning*, 2021.

[14] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[15] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *IEEE ASRU*, 2021.

[16] Jui Shah, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[17] Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King, "Toward a realistic model of speech processing in the brain with self-supervised learning," 2022.

[18] Aditya R. Vaidya, Shailee Jain, and Alexander G. Huth, "Self-supervised models of audio effectively explain human cortical responses to speech," in *Proceedings of International Conference on Machine Learning*, 2022.

[19] Archiki Prasad and Preethi Jyothi, "How accents confound: Probing for accent information in end-to-end speech recognition systems," in *ACL*, 2020.

[20] Chung-Yi Li, Pei-Chieh Yuan, and Hung-Yi Lee, "What does a network layer hear? analyzing hidden representations of end-to-end ASR through speech synthesis," in *IEEE ICASSP*, 2020.

[21] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[22] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, 2014.

[23] Yi Luo and Nima Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE ICASSP*, 2018.

[24] Yi Luo and Nima Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, 2019.

[25] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE ICASSP*, 2017.

[26] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of International Conference on Machine learning*, 2006.

[27] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu, "W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *IEEE ASRU*, 2021.

[28] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[29] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[30] Pengcheng Guo, Xuankai Chang, Shinji Watanabe, and Lei Xie, "Multi-speaker ASR combining non-autoregressive Conformer CTC and conditional speaker chain," in *Interspeech*, 2021.

[31] Naoyuki Kanda, Guoli Ye, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka, "End-to-end speaker-attributed ASR with transformer," in *Interspeech*, 2021.