

# Multi-Channel Multi-Speaker ASR Using Target Speaker’s Solo Segment

Yiwen Shao<sup>1</sup>, Shi-Xiong Zhang<sup>2\*</sup>, Yong Xu<sup>2</sup>, Meng Yu<sup>2</sup>, Dong Yu<sup>2</sup>, Daniel Povey<sup>3</sup>, Sanjeev Khudanpur<sup>1</sup>

<sup>1</sup>Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

<sup>2</sup>Tencent AI Lab, Bellevue, WA, USA

<sup>3</sup>Xiaomi Corp., Beijing, China

yshao18@jhu.edu, zhangshixiong@gmail.com, lucayongxu@global.tencent.com,  
raymondmyu@global.tencent.com, DongYu@ieee.org, dpovey@gmail.com, khudanpur@jhu.edu

Addressing these limitations, our study proposes leveraging

## Abstract

In the field of multi-channel, multi-speaker Automatic Speech Recognition (ASR), the task of discerning and accurately transcribing a target speaker’s speech within background noise remains a formidable challenge. Traditional approaches often rely on microphone array configurations and the information of the target speaker’s location or voiceprint. This study introduces the Solo Spatial Feature (Solo-SF), an innovative method that utilizes a target speaker’s isolated speech segment to enhance ASR performance, thereby circumventing the need for conventional inputs like microphone array layouts. We explore effective strategies for selecting optimal solo segments, a crucial aspect for Solo-SF’s success. Through evaluations conducted on the AliMeeting dataset and AISHELL-1 simulations, Solo-SF demonstrates superior performance over existing techniques, significantly lowering Character Error Rates (CER) in various test conditions. Our findings highlight Solo-SF’s potential as an effective solution for addressing the complexities of multi-channel, multi-speaker ASR tasks.

**Index Terms:** multi-channel multi-speaker ASR, spatial feature, speaker solo segment

## 1. Introduction

Recent advancements in speech processing techniques and deep learning have led to substantial improvements across various automatic speech recognition (ASR) benchmarks [1, 2, 3, 4]. However, accurately recognizing multi-channel, multi-speaker overlapped speech remains challenging, largely due to interfering speakers and background noise [5, 6]. In this complex ASR landscape, the utilization of high-quality, discriminative input features beyond traditional spectral features is crucial for isolating target speech from mixtures. Spatial features, leveraging the phase difference across microphone channels caused by the distinct locations of audio sources, have gained considerable attention [7]. These features form the foundation of many state-of-the-art speech separation [8, 9] and recognition systems [10, 11].

Extending this line of inquiry, Shao et al. introduced a novel perspective by utilizing the room impulse response (RIR) from the target speaker to the microphone array, enhancing spatial feature extraction to include RIR-based spatial features (RIR-SF) [12]. Although RIR-SF approaches an ideal solution with accessible ground truth RIR, the practical acquisition of accurate RIR poses significant challenges, affecting the robustness of RIR-SF in real-world applications.

a short solo segment from the target speaker as an innovative proxy for the actual RIR. This Solo-SF method, through convolution with the overlapped speech signal, aims to overcome the challenges of direct RIR usage by harnessing the unique vocal characteristics of the target speaker. Notably, our approach eliminates the dependency on microphone topology and vision-based positional inputs, facilitating its future application as a universal encoder [13] for diverse multi-channel data setups. Furthermore, we delve into strategies for selecting optimal solo segments and assess our method’s efficacy on both the simulated AISHELL-1 [14] dataset and the real-world AliMeeting dataset [15], demonstrating significant performance enhancements.

## 2. Preliminary: Spatial Feature

Spatial Feature (SF) or Angle Feature (AF), originally introduced by Chen et al. [7], are utilized to underscore the prominence of the target sound source within multi-speaker Time-Frequency (T-F) bins. Unlike most spectral features that depend on the magnitude of the complex Short-Time Fourier Transform (STFT) coefficients  $Y \in \mathbb{C}^{T \times F \times M}$ , where  $T$ ,  $F$ , and  $M$  represent the time, frequency, and channel dimensions respectively, SF is specifically designed to be phase-sensitive. Spatial feature, denoted as  $SF \in \mathbb{R}^{T \times F}$ , leverages the phase differences across channels, attributed to the disparate spatial locations of sound sources. This phase-sensitive design enables SF to effectively differentiate between sources based on their unique spatial characteristics.

### 2.1. 3D spatial feature

Given a pair of microphones, denoted as  $p = (m_1, m_2)$ , within a microphone array, and the multi-channel STFT of the input speech signal, represented as  $Y \in \mathbb{C}^{T \times F \times M}$ , two types of interchannel phase differences are defined: the target-independent interchannel phase difference (IPD) and the target-dependent interchannel phase difference (TPD) as follows:

$$\text{IPD}_{t,f,(p)} = \angle Y_{t,f,m_1} - \angle Y_{t,f,m_2} \quad (1)$$

$$\text{TPD}_{t,f,(p)}(\theta_a, \theta_e, d_o) = \frac{2\pi f}{c(F-1)} \cdot f_s \cdot (d_{m_1} - d_{m_2}) \quad (2)$$

$$d_{m_i} = \sqrt{d_{om_i}^2 + d_o^2 - 2d_{om_i}d_o \cos \theta_a \cos \theta_e} \quad \forall i \in \{1, 2\}$$

where  $\angle$  denotes the phase of a complex number,  $f_s$  is the sampling rate,  $c$  is the speed of sound, and  $F$  is the total number of frequency bands.  $\theta_a$  and  $\theta_e$  represent the azimuth and elevation angles, respectively.  $d_o$ ,  $d_{m_1}$ , and  $d_{m_2}$  are the distances between the target speaker and the microphone(or camera), and from the  $m_1$ -th and  $m_2$ -th microphone to the target speaker,

\*This work was done while Shi-Xiong was at Tencent AI Lab, USA.

respectively. These distances need to be measured using additional visual devices, such as a depth camera.

In [12], Shao et al. propose interpreting the 3D spatial feature (3D-SF) from an alternative perspective of a multiplicative transfer function (MTF) approximation, as described in [16]. When the STFT of the room impulse response (RIR) from the target source's position to the microphone array is available, denoted as  $R \in \mathbb{C}^{K \times F \times M}$ , where  $K$  represents the total length of RIR considered, TPD can also be formulated as:

$$\text{TPD}_{t,f,(p)} = \angle R_{0,f,m_1} - \angle R_{0,f,m_2} \quad (3)$$

where  $\angle R_{0,f,m}$  denotes the phase of the direct wave's RIR from the target source to microphone  $m$ .

Intuitively, the more similar the IPD and TPD are, the higher the likelihood that the mixed time-frequency (T-F) bin is dominated by the target source. By computing the cosine of the difference between IPD and TPD, the 3D spatial feature (3D-SF) can be obtained as:

$$3\text{D-SF}_{t,f} = \cos(\text{IPD}_{t,f,(p)} - \text{TPD}_{t,f,(p)}) \quad (4)$$

## 2.2. RIR Spatial Feature

RIR-SF, introduced in [12], extends the 3D-SF concept by accounting for the impact of reverberant waves on phase differences. It mitigates this effect by convolving the multi-channel speech signal  $Y \in \mathbb{C}^{T \times F \times M}$  with the complex conjugate (H) of the target RIR, denoted as  $R^H \in \mathbb{C}^{K \times F \times M}$ , across the time index. This process defines an intermediate phase known as the RIR-convolved phase (RP):

$$\text{RP}_{t,f,m} = \angle (Y * R^H)_{t,f,m} \quad (5)$$

$$= \angle \left( \sum_{k=0}^{K-1} Y_{t-k,f,m} \cdot R_{k,f,m}^H \right) \quad (6)$$

where  $K$  denotes the length of the RIR considered. In alignment with [12],  $K$  is set to 10 through this work, corresponding to a duration of 0.1 seconds, considering a shift size of 10 ms in the STFT. This practice will be adopted in our work to ensure consistency and comparability with established methodologies.

If  $Y_{t,f,m}$  is dominated by the target source, its phase pattern should align with  $R_{t,f,m}$ , making  $\text{RP}_{t,f,m}$  independent of the microphone and position. The difference across channels should then approach zero. Analogous to the definition of 3D-SF, RIR-SF is quantified as the cosine similarity of the inter-channel RIR-convolved phase differences:

$$\text{RIR-SF}_{t,f} = \cos(\text{RP}_{t,f,m_1} - \text{RP}_{t,f,m_2}) \quad (7)$$

## 3. Proposed: Solo Spatial Feature

### 3.1. Convolution with Solo Segment Instead of RIR

While RIR-SF significantly advances spatial feature extraction by considering reverberant effects, its practicality is limited by the difficulty in accessing accurate ground truth RIRs, making it less effective in unpredictable acoustic environments. Acknowledging the practical challenges of acquiring accurate ground truth RIR, our study proposes a novel solution: the Solo Spatial Feature (Solo-SF). By convolving a selected segment from the target speaker's solo part with the mixed speech signal, Solo-SF sidesteps the reliance on precise RIR data.

To ensure the efficacy of the Solo-SF method, it is imperative that the solo segment  $S \in \mathbb{C}^{K \times F \times M}$  used matches

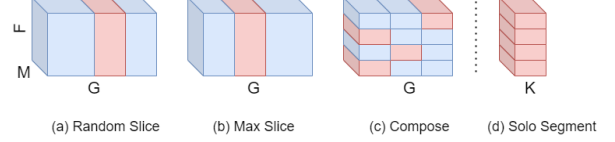


Figure 1: An illustration of 3 different ways of selecting solo segment  $S \in \mathbb{C}^{K \times F \times M}$  from a solo part  $P \in \mathbb{C}^{G \times F \times M}$

the acoustic environment of the longer speech signal  $Y \in \mathbb{C}^{T \times F \times M}$ , specifically **sharing the same underlying RIR**. It can be guaranteed by assuming the target speaker doesn't change its position during conversation. This alignment enables the effective cancellation of phase patterns when  $S^H$ , the conjugate of  $S$ , is convolved with  $Y$ , mirroring the process employed in RIR-SF. We introduce an intermediate phase termed the Solo-convolved phase (SP), defined as:

$$\text{SP}_{t,f,m} = \angle (Y * S^H)_{t,f,m} \quad (8)$$

$$= \angle \left( \sum_{k=0}^{K-1} Y_{t-k,f,m} \cdot S_{k,f,m}^H \right) \quad (9)$$

Subsequently, the Solo-SF is derived as the cosine similarity between interchannel SP differences:

$$\text{Solo-SF}_{t,f} = \cos(\text{SP}_{t,f,m_1} - \text{SP}_{t,f,m_2}) \quad (10)$$

This approach not only mitigates the challenges associated with direct RIR measurements but also capitalizes on the inherent vocal traits of the speaker, facilitating a nuanced and robust spatial feature extraction methodology.

### 3.2. Microphone Array Topology and Position Information Independence

A significant advantage of the proposed Solo-SF method is its independence from microphone array topology and external position information, such as that obtained from vision-based systems, for determining the target speaker's 3D location or estimating the RIR.

This independence marks a departure from previous methodologies that required high-quality, audio-visual aligned data, broadening the applicability of models trained with Solo-SF. Consequently, it allows for seamless adaptation to various microphone array configurations without the need for recalibrating or redesigning the spatial feature extraction process based on specific array geometries or visual input.

On the other hand, obtaining the solo part of the target speaker is notably more feasible in practical applications. This can be achieved through activation by wake words or obtained from on-the-fly or offline diarization systems. This accessibility simplifies the process of acquiring clean solo segments, crucial for the effective application of the Solo-SF method in various real-world scenarios.

There is a notable special case in Equation 9 when  $\|S_{k,f,m}^H\| \approx 0$ , indicating that the solo segment is not producing sound at a particular frequency bin. In such instances,  $S_{k,f,m}^H$  is predominantly influenced by environmental or systematic noise, rendering the solo segment ineffective for distinguishing the desired phase pattern from the mixed speech signal  $Y$ . To tackle this issue, we have devised three methods for selecting  $S \in \mathbb{C}^{K \times F \times M}$  from the target speaker's solo part  $P \in \mathbb{C}^{G \times F \times M}$ .

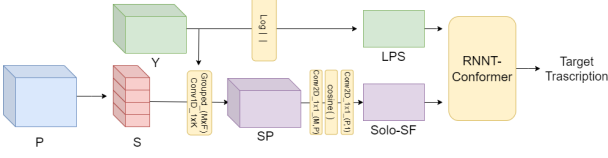


Figure 2: Paradigm of utilizing the target speaker’s solo segment  $S \in \mathbb{C}^{K \times F \times M}$ , selected from a longer solo part  $P \in \mathbb{C}^{G \times F \times M}$ , for target speech recognition in multi-channel, multi-speaker audio  $Y \in \mathbb{C}^{T \times F \times M}$ .

These methods are designed to ensure that the solo segment remains discriminative and useful for phase pattern extraction, even in challenging noise conditions. The approaches and their implementation are illustrated in Figure 1.

1. **Random Slice:** To extract a continuous segment of  $k$  frames from the solo part, we employ a random slicing approach as follows:

$$S_{t,f,m} = P_{c+t,f,m}, \quad c = \text{Rand}[0, G - K] \quad (11)$$

2. **Max Slice:** Select a segment starting from the frame that exhibits the maximum magnitude summation across all frequencies. This process is described as:

$$S_{t,f,m} = P_{c+t,f,m}, \quad c = \arg \max_{t' \in [0, G-K]} \sum_f |P_{t',f,m}| \quad (12)$$

3. **Compose:** In line with Equation 9, since convolution operates on a per-frequency basis, this allows for individual selection of segments for each frequency  $f$ , which are then composited into a new segment. Differing from extracting a uniform segment from the solo part, this method selects  $k$  continuous frames with maximum energy for each frequency  $f$  independently, assembling these into the ultimate solo segment:

$$S_{t,f,m} = P_{c_f+t,f,m}, \quad c_f = \arg \max_{t' \in [0, G-K]} |P_{t',f,m}| \quad (13)$$

## 4. Implementation

Figure 2 illustrates the comprehensive framework we have developed for the extraction of Solo-SF and its application within a speech recognition system. This process initiates with the Short-Time Fourier Transform (STFT) of both the mixed speech signal  $Y$  and the target speaker’s solo part  $P$ , employing a 25 ms window and a 10 ms hop length at a sampling rate of 16 kHz. All components are efficiently implemented as fully differentiable modules within the PyTorch framework, specifically using `nn.Module`.

The conversion of Equation 9 into a computational operation is achieved through the use of `nn.functional.conv1d`, configured with  $F \times M$  groups, and utilizing the target solo segment  $S \in \mathbb{C}^{K \times F \times M}$  as the convolution kernel. To transition from SP to Solo-SF, two `nn.Conv2d` layers with specially designed parameters (i.e. 1’s and -1’s) are employed to compute pairwise interchannel SP differences and their summation, yielding the final Solo-SF. These layers are maintained in a fixed state (frozen) within this study to facilitate future investigations.

Additionally, we extract the logarithmic power spectrum (LPS) from the reference channel, defined as  $\text{LPS}_{t,f} = \log |Y_{t,f,m=1}|$ , which serves as a spectral feature. This is then

Table 1: Character Error Rate (CER) % on AISHELL-1 simulated dev/test sets.  $\dagger$  Ground truth relative positions of microphones and target speakers, with uncertainties of  $\pm 0.5\text{m}$  in their absolute positions within the room. It provides 3D-SF with ground truth  $(\theta_a, \theta_e, d_o)$  and provides RIR-SF with partial information for estimating RIR using image source method (ISM).  $\ddagger$  Indicates the use of Ground Truth RIR (RIR-GT).

Method	CER% (dev/test)	
	RT60=(0.1, 0.6)s	RT60=(0.5, 0.7)s
Single-speaker	8.57/9.71	10.64/11.99
3D-GT $\dagger$	14.11/15.67	20.28/21.26
RIR-EST $\dagger$	22.37/24.33	22.49/24.76
RIR-GT $\ddagger$	<b>10.03/11.28</b>	<b>10.90/12.38</b>
Solo-random	16.05/17.99	21.39/23.28
Solo-max	12.70/14.48	16.80/18.26
Solo-compose	<b>11.57/13.23</b>	<b>13.70/15.56</b>

concatenated with the phase-sensitive Solo-Spatial Feature to form a composite feature vector with dimensions  $[T \times 2F]$ . This composite feature is input into a downstream Conformer-based RNN-T ASR model, as detailed in [17], comprising a 12-layer, 4-head Conformer encoder [18] with 512 attention dimensions and 2048 feed-forward dimensions.

## 5. Experiments

### 5.1. Simulated Data: AISHELL-1

To ensure our proposed method’s comparability with existing approaches, we adopt the data simulation practice outlined in [11, 12] using the AISHELL-1 dataset [14] for our experiments. The Pyroomacoustics toolkit [19], leveraging the image-source method (ISM) [20], serves as the basis for Room Impulse Response (RIR) generation and estimation. The simulation parameters include room size, RT60 (reverberation time), microphone position, and speaker positions. Room dimensions vary between  $[3, 3, 2.5]$  and  $[8, 6, 4]$  meters, featuring one microphone array and two speakers in each scenario. The microphone array consists of an 8-element non-uniform linear array with spacings of 15-10-5-20-5-10-15 cm. RT60 values are randomly chosen from 2 settings, ranging from  $[0.1, 0.6]$  seconds and  $[0.5, 0.7]$  seconds, reflecting typical room conditions of weak reverberation and strong reverberation respectively. Accordingly, 100,000 sets of RIRs are pre-generated for each configuration. During training, multi-channel reverberant overlapped speech signals are synthesized on-the-fly by convolving the pre-generated RIRs with dry, clean speech samples from AISHELL-1, with signal-to-interference ratios (SIRs) randomly selected between  $-6$  and  $6$  dB. The overlap ratio for the two speakers within a mixed utterance varies from 0.5 to 1, ensuring a diverse set of scenarios for model training and evaluation.

**Best Solo Segment Selection Method – Compose:** For an overlapped utterance  $Y$ , the solo part  $P$  is consistently selected as a fixed 2-second long, random, continuous speech from the target before mixing with interfering speech. As illustrated in the last three rows of Table 1, employing the “Compose” (i.e. Equation 12) method for solo segment selection consistently yields the best results across all three tested methods. This outcome supports our initial hypothesis that maximizing coverage of frequency bins within the solo segment is crucial for optimal performance during convolution.

**3D V.S. RIR V.S. Solo:** RIR-SF, when utilizing ground truth RIR, approaches the performance of the single-speaker lower

Table 2: Character Error Rate (CER) % on Alimeeting. † Eval-Rev represents evaluation under reverberation without additional interfering speech. ‡ Eval-Simu corresponds to simulation same as that for AISHELL-1 with RT60=(0.1, 0.6)s.

Method	Training Data	Additional Input	Single-Speaker		Multi-Speaker		
			Eval-Near	Eval-Rev†	Eval-Simu‡	Eval-Far	Test-Far
Single-Channel	Near	None	14.72	29.02	98.08	55.25	56.28
	Far		20.16	23.03	87.27	34.23	37.00
	Near-Rev†+Far		15.82	17.63	83.56	32.13	34.55
3D-GT	Simu‡	Microphone topology, spk-to-mic relative position	N/A	18.96	22.72	N/A	
RIR-EST		Estimated RIR (from mic topo, spk and mic absolute position, room size)		20.94	30.04		
RIR-GT		Ground truth RIR		16.46	18.53		
Solo-compose (proposed)		Simu‡ Far Simu‡+Far		2 seconds nearest solo part from diarization	17.11 24.79 <b>17.59</b>	20.32 52.23 <b>21.08</b>	40.75 29.59 <b>26.83</b>
SOT [21, 15]	Near	None	N/A			30.80	32.40
SOT <sub>bf</sub> [21, 15]	Near+ Far	CDDMA beamformer [22] fixed mic topo				29.70	30.90

bound, highlighting its potential efficacy. However, its performance significantly relies on the accuracy of the RIR information. Even when combining partial correct information with ground truth for 3D-SF, its effectiveness diminishes, resulting in outcomes inferior to those achieved by 3D-SF. Conversely, Solo-SF, which requires only a 2-second snippet of solo speech from the target speaker and no vision-based information, delivers results nearly on par with the ideal yet impractical lower bound set by ground truth RIR-SF. Furthermore, it outperforms 3D-SF, establishing Solo-SF as a more dependable alternative in this comparative analysis.

## 5.2. Real Data: AliMeeting

As discussed in Section 3.2, a distinct advantage of the proposed Solo-SF is its independence from both vision-based inputs and the topology of the microphone array. This attribute significantly enhances its applicability to real-world scenarios and publicly available datasets. For the purpose of our evaluation, we chose the AliMeeting dataset [15], a Mandarin-language corpus collected from real meeting environments and specifically designed for the ICASSP 2022 M2MeT challenge. This dataset includes multi-channel far-field speech recordings, amassing a total of 104.75 hours for training, 4 hours for evaluation, and 10 hours for testing, alongside simultaneously recorded near-field data from the meeting participants. Each recorded session, ranging from 15 to 30 minutes, features discussions among 2-4 participants. The training data presents an overlap ratio of 42.27%, which is significantly lower than that of our previously discussed simulated dataset, yet it offers a closer approximation to authentic meeting situations. Furthermore, the challenge documentation stipulates that participants were required to remain stationary during recordings, a condition that dovetails with the operational prerequisites for the effective application of Solo-SF, thereby ensuring the consistency of the Room Impulse Response (RIR) between the solo segments and the speech intended for transcription.

Given the availability of ground truth diarization data from the M2MeT challenge, extracting the solo parts of the target speaker for analysis becomes straightforward. To ensure the consistency of the underlying RIR between the solo segments and the speech to be transcribed, we meticulously select a 2-second solo segment closest in time to the current utterance. This selection is made irrespective of whether the solo part falls within or outside the current utterance, ensuring optimal relevance and acoustic similarity for Solo-SF application.

In alignment with common practices for leveraging open-source data, incorporating near-field clean speech or its simulations into the training data has been shown to enhance ASR model performance. Adhering to this approach, we enriched our training dataset with two variations using the near-field Train set: one with only reverberation (Near-Rev) and another with both reverberation and interference speech (Simu). Given the unavailability of additional inputs for 3D-SF and RIR-SF within the Alimeeting dataset, we generated two evaluation sets from the near-field Eval set, termed Eval-Rev and Eval-Simu, facilitating a direct comparison of Solo-SF with 3D-SF and RIR-SF. As evidenced by the results in Table 2, and consistent with observations from AISHELL-1, Solo-SF outperforms its counterparts by demonstrating superior overall performance and robustness across all evaluated spatial features.

Focusing on the real test scenarios, Eval-Far and Test-Far, Solo-SF, when trained on a combination of simulated and far-field data, significantly outperforms the single-channel baseline system—achieving a substantial reduction in CER by absolute margins of 5.30% and 5.03%, respectively. Additionally, when compared to the official baseline SOT and its beamformer-enhanced variant, Solo-SF exhibits superior performance without necessitating additional modules. This underscores its efficacy and potential applicability in multi-channel, multi-speaker speech recognition tasks, demonstrating its capability to deliver enhanced speech recognition accuracy in complex acoustic environments.

## 6. Conclusion

In this work, we have advanced the field of multi-channel, multi-speaker ASR by presenting a comprehensive study on the utilization of a target speaker’s solo segment. Central to our approach is the introduction of Solo-SF, a novel spatial feature extraction method designed to enhance ASR performance without relying on traditional inputs such as microphone array topology or vision-based data. Our investigation further delves into optimal strategies for selecting solo segments, a critical component for ensuring the effectiveness of Solo-SF in diverse acoustic environments.

Our findings highlight the efficacy of using solo segments for enhancing ASR systems, suggesting avenues for future research in optimizing segment selection. The Solo-SF approach sets a promising direction for improving speech recognition accuracy and robustness in real-world environments.

## 7. References

- [1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [2] Y. Wang, T. Chen, H. Xu, S. Ding, H. Lv, Y. Shao, N. Peng, L. Xie, S. Watanabe, and S. Khudanpur, “Espresso: A fast end-to-end neural speech recognition toolkit,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 136–143.
- [3] Y. Shao, Y. Wang, D. Povey, and S. Khudanpur, “Pychain: A fully parallelized pytorch implementation of lf-mm for end-to-end asr,” *arXiv preprint arXiv:2005.09824*, 2020.
- [4] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, “Pushing the limits of semi-supervised learning for automatic speech recognition,” *arXiv preprint arXiv:2010.10504*, 2020.
- [5] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, “Far-field automatic speech recognition,” *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2020.
- [6] Y. Masuyama, X. Chang, S. Cornell, S. Watanabe, and N. Ono, “End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 260–265.
- [7] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, “Multi-channel overlapped speech recognition with location guided speech extraction network,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.
- [8] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, “Multi-modal multi-channel target speech separation,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020.
- [9] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, “Adl-mvdr: All deep learning mvdr beamformer for target speech separation,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6089–6093.
- [10] J. Yu, S.-X. Zhang, B. Wu, S. Liu, S. Hu, X. Liu, H. M. Meng, and D. Yu, “Audio-visual multi-channel integration and recognition of overlapped speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.
- [11] Y. Shao, S.-X. Zhang, and D. Yu, “Multi-channel multi-speaker asr using 3d spatial feature,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6067–6071.
- [12] —, “Rir-sf: Room impulse response based spatial feature for multi-channel multi-talker asr,” *arXiv preprint arXiv:2311.00146*, 2023.
- [13] Z. Huang, Y. Shao, S.-X. Zhang, and D. Yu, “Unix-encoder: A universal  $x$ -channel speech encoder for ad-hoc microphone array speech processing,” *arXiv preprint arXiv:2310.16367*, 2023.
- [14] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.
- [15] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma *et al.*, “M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6167–6171.
- [16] Y. Avargel and I. Cohen, “On multiplicative transfer function approximation in the short-time fourier transform domain,” *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.
- [17] F. Kuang, L. Guo, W. Kang, L. Lin, M. Luo, Z. Yao, and D. Povey, “Pruned rnn-t for fast, memory-efficient asr training,” *arXiv preprint arXiv:2206.13236*, 2022.
- [18] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.
- [19] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: A python package for audio room simulation and array processing algorithms,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 351–355.
- [20] E. A. Lehmann and A. M. Johansson, “Prediction of energy decay in room impulse responses simulated with an image-source model,” *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.
- [21] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, “Serialized output training for end-to-end overlapped speech recognition,” *arXiv preprint arXiv:2003.12687*, 2020.
- [22] W. Huang and J. Feng, “Differential beamforming for uniform circular array with directional microphones,” in *Interspeech*, 2020, pp. 71–75.