



# REAL-T: Real Conversational Mixtures for Target Speaker Extraction

Shaole Li<sup>1</sup>, Shuai Wang<sup>2,1,3,\*</sup>, Jiangyu Han<sup>4</sup>, Ke Zhang<sup>1,3</sup>, Wupeng Wang<sup>5</sup>, Haizhou Li<sup>1,3,5</sup>

<sup>1</sup>Shenzhen Research Institute of Big Data, Shenzhen, China

<sup>2</sup>School of Intelligence Science and Technology, Nanjing University, Suzhou, China

<sup>3</sup>School of Data Science, The Chinese University of Hong Kong, Shenzhen, China

<sup>4</sup>Brno University of Technology, Speech@FIT, Czechia

<sup>5</sup>Department of Electrical and Computer Engineering, National University of Singapore, Singapore

shuaiwang@nju.edu.cn

## Abstract

Current target speaker extraction (TSE) systems achieve remarkable performance on synthetic datasets like LibriMix and WSJMix. However, their effectiveness in real conversational scenarios, where the cocktail party problem is most prevalent, remains largely unexplored. In this paper, we conduct a comprehensive analysis of several speaker diarization datasets and introduce REAL-T, the first conversation-centric dataset specifically designed for TSE in real-world conditions. Our evaluations reveal significant performance degradation of existing TSE models on this dataset, highlighting the unaddressed complexity of real-world speech extraction. To facilitate controlled benchmarking, we define two subsets: BASE and PRIMARY, ensuring more manageable yet challenging evaluation settings. **Index Terms:** Real-world, conversational, target speaker extraction, dataset, REAL-T

## 1. Introduction

While datasets like LibriMix [1] and WSJMix [2] are standard for speech separation (SS) [3, 4] and target speaker extraction (TSE) [5–10], their synthetic construction introduces key deviations from the real-world. Unlike natural recordings, where concurrent voices share the similar acoustic environment, simulated mixtures combine utterances under different conditions, disrupting loudness relationships. Additionally, the lack of real reverberation and ambient noise results in an unnatural acoustic signature. More critically, these datasets primarily feature read speech from fixed texts, missing the spontaneous interactions and reactive overlaps that characterize real conversational settings and pose key challenges in the cocktail party problem.

To alleviate the above limitation, some works like REAL-M [11] and LibriCSS [12] propose to simulate natural speech overlaps through controlled recording setups: REAL-M achieves this by asking multiple speakers reading different LibriSpeech texts in shared acoustic environments; LibriCSS involves simultaneous playback and re-recording of pre-existing isolated utterances. However, both datasets still cannot fully capture the characteristics of real conversations. For example, meeting scenarios typically involve speakers engaging intermittently over extended durations, with sporadic utterances, varying speaker turns, and natural speech overlaps. Moreover, these two datasets do not provide separate audios for target speaker registration, making it unsuitable for direct use in TSE research. Conversely, speaker diarization datasets naturally capture real conversational dynamics, providing both overlapping speech and sufficient non-overlapping segments that can be utilized for target speaker registration.

\*Corresponding author

Building on these insights, we analyzed existing speaker diarization datasets and developed an automated filtering pipeline to segment data while preserving semantic completeness, generating realistic mixtures and the corresponding enrollment utterances. The obtained REAL-T is the first conversational-centric TSE dataset, characterized by the following key features:

- **Multi-lingual:** Mandarin and English data are included.
- **Multi-Genre:** It covers diverse scenarios and styles.
- **Multi-Enrollment:** Each target speaker is associated with multiple enrollment utterances for improved generalization.

During the construction process of our dataset, we identified several real-world challenges that are rarely considered in the existing simulated data, including:

- Most speakers overlap infrequently, and fully overlapping cases are rare.
- Some speakers exhibit frequent but non-continuous speaking patterns, increasing the complexity of the task.
- Conversational speech, especially in meetings, often consists of short-term segments, making extraction harder.
- Environmental noise and non-speech vocalizations, such as laughter and humming, further complicate speech extraction.

Due to the complex of real-world scenarios, we observed that the selected systems often perform extremely bad. To facilitate meaningful evaluation, we filtered out outlier samples and created **BASE** and **PRIMARY** test sets from a more manageable subset of the original collected data. Additionally, we provide comprehensive metadata for the full dataset, including information such as gender and enrollment details, enabling researchers to conduct more detailed analyses and investigations<sup>1</sup>.

## 2. Dataset

### 2.1. Corpus selection

Although various real-world conversational datasets [13–22] are available for speaker diarization, constructing a publicly accessible TSE dataset from these resources is not straightforward. These challenges stem from data licensing restrictions, sparse speaker interactions, and the lack of high-quality transcriptions. Our aim is to develop a diverse TSE dataset that ensures high-quality ASR transcriptions. To this end, we selected public available datasets including AISHELL-4 [13], AI-Meeting [14, 15], AMI [16], CHiME6 [17], and DipCo [18], utilizing only their test set. Detailed information of these selected datasets is shown in Table 1, where the overlap percentages (ovl) are computed using the diarization\_utils toolkit<sup>2</sup>.

<sup>1</sup><https://real-tse.github.io>

<sup>2</sup>[https://github.com/BUTSpeechFIT/diarization\\_utils/blob/main/compute\\_ratios.py](https://github.com/BUTSpeechFIT/diarization_utils/blob/main/compute_ratios.py)

Table 1: Information of corpus for constructing REAL-T.

Corpus	#Files	#Spk	#Hours	Ovl (%)	Characteristics
AliMeeting	20	2-4	10.8	20.36	Meeting, Mandarin
AISHELL-4	20	5-7	12.7	4.95	Meeting, Mandarin
AMI	16	3-4	9.1	14.58	Meeting, English
DipCo	5	4	2.6	27.48	Dinner, English
CHiME6	2	4	5.2	33.92	Dinner, English

## 2.2. Data preprocessing

### 2.2.1. Audio

To capture overlapping speech, we use the first microphone channel for AISHELL-4, AliMeeting, and AMI. For CHiME-6 and DipCo, which feature spatially distributed microphone arrays, we average all channels to produce a single-channel recording that integrates conversations from different regions.

### 2.2.2. Transcriptions

Original transcripts often contain various tags (e.g. [noise], [laugh], [inaudible], [redacted], etc.) and extraneous punctuation (e.g., !, ?, etc.), which introduce challenges when computing ASR metrics. Therefore, we employ Whisper-large-v2 normalizer to further standardize the transcripts. The final processed annotations are structured in a format akin to RTTM while enriching with detailed transcript information.

## 2.3. Construction of TSE data

### 2.3.1. Mixture utterances

To ensure semantic completeness, we generate mixture utterances directly from the original long recordings, focusing only on overlapping segments with a cumulative overlap duration of at least 5 seconds. The extraction process is as follows:

1. *Sort*: Arrange segments in ascending order of start time.
2. *Check for Overlaps*: Iterate through the sorted segments and detect overlaps using timestamps.
3. *Merge Overlapping Segments*: If overlaps are detected, merge the segments into a single, semantically complete one and re-check for further overlaps (Back to Step 2).
4. *Calculation*: If no further overlaps are found, mark as a new mixture utterance and compute the statistics such as overlap duration, start time, end time, and total duration.

The distribution of mixture utterance duration for our selected results is shown in Figure 1. Most utterances are under 50 seconds, with few exceeding 100 seconds. To maintain consistency, these outliers were excluded from the final dataset.

### 2.3.2. Enrollment utterances

Enrollment utterances are extracted from the original non-overlapping segments. To ensure sufficient speech content for reliable speaker enrollment, each selected utterance is at least 5 seconds long. To capture enrollment variability and to prevent imbalance, up to five utterances per speaker are randomly selected. If fewer than five are available, all will be used.

## 3. Experiments

### 3.1. Configuration

We first use an open-source TSE system to validate the constructed dataset. Specifically, we employ the BSRNN-based

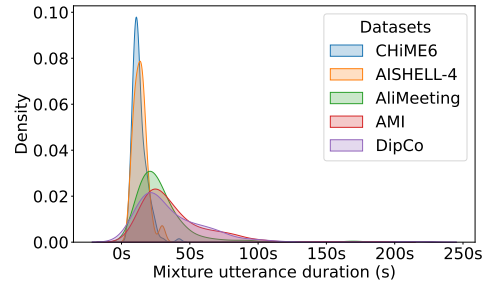


Figure 1: Probability density distribution of mixture utterances duration across different datasets.

TSE model from WeSep<sup>3</sup> [23], trained on the VoxCeleb1 dataset [24]. Since the ground truth for clean sources is unavailable, we evaluate ASR performance on the extracted speech using token error rate (TER), where the token refers to Mandarin characters and English words, respectively. For English, we utilize Whisper-large-v2<sup>4</sup> [25], and for Mandarin, FireRedASR-AED-L<sup>5</sup> [26]. All transcriptions are normalized using the normalizer provided by Whisper-large-v2.

### 3.2. BASE Test Set

The BASE test set is a filtered dataset designed for robust ASR evaluation. To define selection criteria, we analyzed ASR performance on speech extracted by the TSE system, focusing on the relationship between TER and the target speaker’s proportion in the mixture utterance.

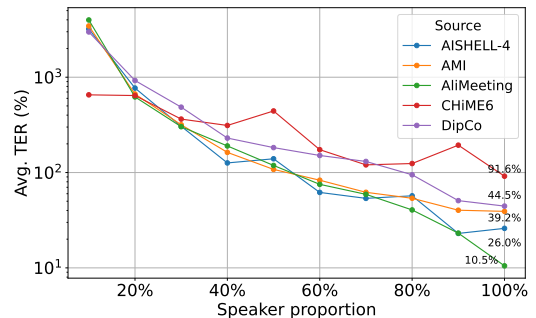


Figure 2: Avg. TER vs. speaker proportion.

As shown in Figure 2, we divided the speaker proportion into ten intervals (0-10%, 10-20%, ..., 90-100%) and found that ASR performance improves significantly as the target speaker’s proportion increases. However, performance is extremely poor when the proportion is below 20%. Additionally, some target speakers did not produce actual words but instead laughed or hummed. To ensure meaningful speech content, we included only target speakers with a proportion above 20% and further filtered out those whose ground truth transcript contained five or fewer English words or Chinese characters. It’s also worth noting from Table 2 that BSRNN-based TSE model’s performance on BASE test set is extremely poor. This high difficulty makes it challenging to effectively evaluate the TSE model. Therefore, based on the analysis of the BASE Test Set, we propose constructing a new, more suitable subset for effective evaluation.

<sup>3</sup><https://github.com/wenet-e2e/WeSep>

<sup>4</sup><https://huggingface.co/openai/whisper-large-v2>

<sup>5</sup><https://github.com/FireRedTeam/FireRedASR>

Table 2: Avg. TER (%) on BASE and PRIMARY test sets

Language	Source	BASE	PRIMARY
Chinese	AISHELL-4	96.37	40.87
	AliMeeting	117.25	65.97
	Overall	109.67	57.61
English	AMI	104.30	50.33
	CHiME6	145.37	92.46
	DipCo	185.37	61.97
	Overall	124.25	69.63

### 3.3. PRIMARY Test Set

Furthermore, we create a PRIMARY Test set by only considering the primary speaker (i.e., the speaker with the maximum duration proportion) in each mixture utterance of the BASE test set, where the duration of each utterance is limited to 30 seconds. We take it as the main test set for current REAL-T<sup>6</sup>.

#### 3.3.1. Statistics of PRIMARY

The PRIMARY subset statistics are summarized in Table 3, with a total mixture duration of 156.03 minutes and a 70.26 minute overlap. The distributions of mixture ratio and speaker proportion are shown in Figure 3.

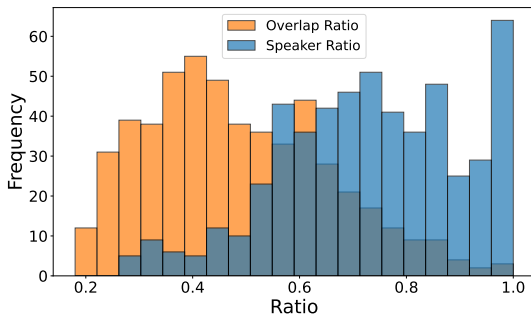


Figure 3: Distribution of overlap ratio and speaker proportion

#### 3.3.2. Impact of number of speakers in mixture utterance

Table 4 presents statistics on the number of speakers in mixtures within the PRIMARY test set. As the number of speakers in a mixture utterance increases, the difficulty of extraction also rises, with the 5-speaker scenario being an exception due to having only a single utterance. It is worth noting, however, that the data for the 2-speaker is also relatively scarce, with the majority of mixtures in the PRIMARY test set focusing on 3-4 speaker scenarios, reflecting the challenging nature of this test set in real-world environments.

#### 3.3.3. Impact of different enrollment utterances

Figure 4 Box plot distribution of TER standard deviation across different enrollment utterances for the same speaker, categorized by source. It is evident that the choice of enrollment utterance significantly influences the extraction performance, especially in AISHELL-4, CHiME6, DipCo datasets. To investigate the large variance from the perspective of speaker embeddings, we computed the similarity between different enrollment utterances using the WeSpeaker toolkit [27, 28]<sup>7</sup>. For example, in the case of the mixture utterance

<sup>6</sup>The current version is designed at a moderate level of difficulty

<sup>7</sup><https://github.com/wenet-e2e/wespeaker>

S\_R003S04C01\_mixture\_1176.25.1185.92, Its target speaker 005-M. Among the five enrollment utterances, the best-performing ones resulted in a TER of 44.44%, while the worst led to 129.63%. This difference could be attributed to the higher background noise in the poorer-quality enrollment utterances. Furthermore, the speaker embeddings from the better-performing utterances showed about 50% similarity to those associated with the poor extraction results.

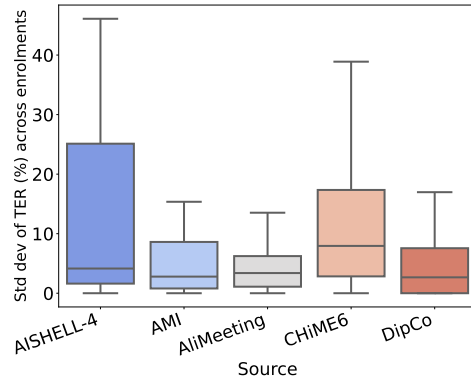


Figure 4: Distribution of the standard deviation of TER (%) across different enrolments for the same target speaker

#### 3.3.4. Evaluation of existing methods

We then evaluated the performance of several recently proposed TSE models on the REAL-T-PRIMARY test set, including BSRNN [29], BSRNN\_HR [30], USEF-TFGridnet [31], and TSELM-L [32] models, which were pretrained on different datasets. All systems are re-implemented using WeSep [23] except for TSELM-L, for which we directly use the official checkpoint<sup>8</sup>. The results of the evaluation are shown in Table 5.

Interestingly, the TSELM-L model, which is based on a generative approach, exhibits quite poor results, especially on the Chinese portion. We checked some of the extracted samples and found that even when the mixture and enrollment speech are Chinese, the extracted speech sounds like English pronunciation but is not interpretable. This phenomenon, to some extent, reflects the strong language dependency of generative methods.

For the BSRNN and BSRNN\_HR models trained on VoxCeleb1, we observed that their performance was comparable on the simulated dataset Libri2Mix-360, but significantly better on REAL-T—a real-world, out-of-domain dataset—compared to models trained on Libri2Mix. This aligns with the results presented in [23]. Additionally, when comparing the two models trained on Libri2Mix-100, Libri2Mix-360 and VoxCeleb1, namely BSRNN and BSRNN\_HR, we found that BSRNN\_HR generally outperformed BSRNN.

Although USEF-TFGridnet<sup>9</sup> performed well on simulated datasets, its performance on the real-world REAL-T dataset was notably poor, indicating overfitting to Libri2Mix-100. From Table 2, it is evident that the performance on CHiME6 and DipCo datasets is the worst, both in the BASE and PRIMARY test sets. This is likely due to the Dinner Party scenario being more challenging than meeting scenarios, with more noise, complex environments, and multi-room conditions, which make extraction more difficult.

<sup>8</sup>[https://www.modelscope.cn/datasets/wenet/wesep\\_pretrained\\_models/](https://www.modelscope.cn/datasets/wenet/wesep_pretrained_models/)

<sup>9</sup>The training for this model is extremely slow, thus we only show the results on Libri2Mix-100

Table 3: Statistics for the PRIMARY test set

Lang: Language (zh for Chinese, en for English); T. Dur (min): Total mixture utterance duration (minutes); Ovl Dur (min): Total overlap duration in mixture utterances (minutes); # Utt: Total number of mixture utterances; Avg. Ovl. Ratio: Average overlap ratio per mixture utterance; # Test: Total number of test samples.

Category	Source	Lang	T. Dur (min)	Ovl Dur (min)	# Utt	Avg. Ovl. Ratio	# Test
<b>By source</b>							
	AISHELL-4	zh	10.37	5.18	46	0.53	240
	AliMeeting	zh	52.64	22.31	162	0.45	481
	AMI	en	42.22	17.15	122	0.42	592
	CHiME6	en	26.67	15.44	123	0.61	545
	DipCo	en	24.13	10.18	75	0.44	133
<b>By language</b>							
Overall	Total	-	156.03	70.26	528	0.49	1991
	English (en)	en	93.02	42.77	320	0.50	1270
	Chinese (zh)	zh	63.01	27.49	208	0.47	721

Table 4: Avg. TER (%) & Duration by Lang. & Speaker Count

Spk #	Lang	Avg. (%)	Dur (min)	Ovl (min)
2	en	43.47	5.37	1.99
	zh	27.75	3.11	1.27
3	en	58.22	30.57	13.12
	zh	42.57	15.67	6.80
4	en	78.69	57.07	27.66
	zh	73.28	44.00	19.32
5	zh	32.56	0.23	0.1

Table 5: Comparison of model performance on the simulated Libri2Mix and PRIMARY test set. The best TERs are in bold.

Model	Training data	Libri2Mix	PRIMARY Test set	
		SI-SDR(dB)	zh(%)	en(%)
TSELM-L	Libri2Mix-360	/	331.73	192.39
USEF-TFGridnet	Libri2Mix-100	<b>18.05</b>	67.98	87.27
BSRNN	Libri2Mix-100	12.95	81.74	91.20
	Libri2Mix-360	16.57	69.80	73.61
	VoxCeleb1	16.50	<b>57.61</b>	69.63
BSRNN_HR	Libri2Mix-100	15.91	70.03	78.96
	Libri2Mix-360	17.99	63.38	74.64
	VoxCeleb1	16.38	58.77	<b>66.46</b>

### 3.3.5. Example analysis

The dataset we constructed poses significant challenges, as it includes real-world scenarios with substantial overlap between speakers. For instance, in the PRIMARY test set, the Mixture segment EN2002a\_mixture\_0.00\_25.26 was processed using BSRNN, yielding average speaker extraction accuracies of 311.11%, 201.19%, 31.58%, and 89.70% for speakers FEO070, MEE071, MEE073, and FEO072, respectively. Despite the presence of two male and two female speakers, only one male and one female speaker were effectively extracted. Among them, MEE073 demonstrated the best performance, while FEO072 exhibited suboptimal results. As illustrated in Figure 5, although MEE071 is the primary speaker, its extraction was predominantly comprised of MEE073’s speech due to the latter’s higher initial proportion. Furthermore, FEO070 had the lowest speaker ratio, while FEO072—despite having a

higher overall proportion—had most of its speech concentrated toward the end of the mixture. Both cases presented significant challenges for extraction. These findings underscore the challenging of real-world TSE, where performance often drops significantly compared to simulated data.

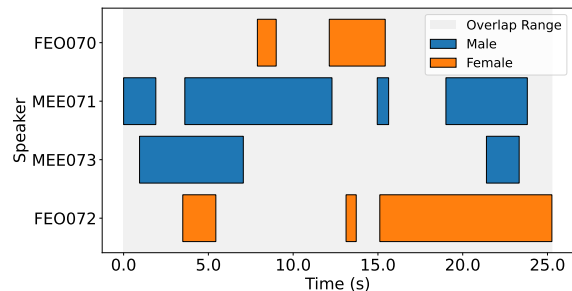


Figure 5: Example of mixture utterance

## 4. Conclusion

In this study, we introduced REAL-T, a real conversational dataset for target speaker extraction, constructed from the well-known speaker diarization datasets. We defined BASE and PRIMARY test sets to balance performance and complexity. Experiments on REAL-T-PRIMARY reveal significant performance degradation in existing TSE models, highlighting the limitations of the existing simulated data. To advance related research, all datasets, benchmarks, and metadata will be open-sourced.

## 5. Acknowledgement

This work was supported by National Natural Science Foundation of China, (Grant No. 62401377), Shenzhen Science and Technology Program (Shenzhen Key Laboratory, Grant No. ZDSYS20230626091302006), Shenzhen Science and Technology Research Fund (Fundamental Research Key Project, Grant No. JCYJ20220818103001002), Program for Guangdong Introducing Innovative and Entrepreneurial Teams, (Grant No. 2023ZT10X044), Yangtze River Delta Science and Technology Innovation Community Joint Research Project (Grant No. 2024CSJGG01100).

## 6. References

- [1] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.
- [2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 31–35.
- [3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [4] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [5] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.
- [6] K. Liu, Z. Du, X. Wan, and H. Zhou, "X-sepformer: End-to-end speaker extraction network with explicit optimization on speaker confusion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [7] F. Hao, X. Li, and C. Zheng, "X-tf-gridnet: A time-frequency domain target speaker extraction network with adaptive speaker embedding fusion," *Information Fusion*, vol. 112, p. 102550, 2024.
- [8] X. Yang, C. Bao, J. Zhou, and X. Chen, "Target speaker extraction by directly exploiting contextual information in the time-frequency domain," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 476–10 480.
- [9] J. Peng, M. Delcroix, T. Ochiai, O. Plchot, S. Araki, and J. Černocký, "Target speech extraction with pre-trained self-supervised learning models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 421–10 425.
- [10] J. Li, K. Zhang, S. Wang, H. Li, M.-W. Mak, and K. A. Lee, "On the effectiveness of enrollment speech augmentation for target speaker extraction," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 325–332.
- [11] C. Subakan, M. Ravanelli, S. Cornell, and F. Grondin, "Real-m: Towards speech separation on real mixtures," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6862–6866.
- [12] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7284–7288.
- [13] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu *et al.*, "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," *arXiv preprint arXiv:2104.03603*, 2021.
- [14] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma, X. Xu, and H. Bu, "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *Proc. ICASSP*. IEEE, 2022.
- [15] F. Yu, S. Zhang, P. Guo, Y. Fu, Z. Du, S. Zheng, W. Huang, L. Xie, Z.-H. Tan, D. Wang, Y. Qian, K. A. Lee, Z. Yan, B. Ma, X. Xu, and H. Bu, "Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge," in *Proc. ICASSP*. IEEE, 2022.
- [16] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The ami meeting corpus."
- [17] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.
- [18] M. Van Segbroeck, A. Zaid, K. Kutsenko, C. Huerta, T. Nguyen, X. Luo, B. Hoffmeister, J. Trmal, M. Omologo, and R. Maas, "Dipco-dinner party corpus," 2020.
- [19] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First dihard challenge evaluation plan," in *tech. Rep.* Linguistic Data Consortium, University of Pennsylvania, 2018.
- [20] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, and M. L. Ganapathy, "The second dihard diarization challenge: Dataset, task, and baselines," 2019.
- [21] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, J. D. Cieri, S. Ganapathy, and M. Liberman, "The third dihard diarization challenge," 2021.
- [22] J. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: Speaker diarisation in the wild," 2020.
- [23] S. Wang, K. Zhang, S. Lin, J. Li, X. Wang, M. Ge, J. Yu, Y. Qian, and H. Li, "Wesep: A scalable and flexible toolkit towards generalizable target speaker extraction," in *Interspeech 2024*, 2024, pp. 4273–4277.
- [24] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.
- [25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [26] K.-T. Xu, F.-L. Xie, X. Tang, and Y. Hu, "Firedasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration," *arXiv preprint arXiv:2501.14350*, 2025.
- [27] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [28] S. Wang, Z. Chen, B. Han, H. Wang, C. Liang, B. Zhang, X. Xiang, W. Ding, J. Rohdin, A. Silnova *et al.*, "Advancing speaker embedding learning: Wespeaker toolkit for research and production," *Speech Communication*, vol. 162, p. 103104, 2024.
- [29] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.
- [30] K. Zhang, J. Li, S. Wang, Y. Wei, Y. Wang, Y. Wang, and H. Li, "Multi-level speaker representation for target speaker extraction," *arXiv preprint arXiv:2410.16059*, 2024.
- [31] B. Zeng and M. Li, "Usef-tse: Universal speaker embedding free target speaker extraction," *arXiv preprint arXiv:2409.02615*, 2024.
- [32] B. Tang, B. Zeng, and M. Li, "Tselm: Target speaker extraction using discrete tokens and language models," *arXiv preprint arXiv:2409.07841*, 2024.