REAL-T: Real Conversational Mixtures for Target Speaker Extraction

Addressing the Gap Between Synthetic and Real-World TSE Performance

Shaole Li, **Shuai Wang***, Jiangyu Han, Ke Zhang, Wupeng Wang, Haizhou Li

Nanjing University
Chinese University of Hong Kong (Shenzhen)
Brno University of Technology
National University of Singapore

August 19, 2025

Outline

The Challenge: Synthetic vs. Real-World TSE

Current State:

- TSE systems excel on synthetic datasets (LibriMix, WSJMix)
- Remarkable performance in controlled environments
- Standard benchmarks for speech separation research

The Problem:

- Performance gap in real conversational scenarios
- Cocktail party problem remains unsolved
- Limited evaluation on authentic speech overlaps

Key Limitation

Synthetic datasets fail to capture real-world acoustic complexity, spontaneous interactions, and natural speech overlaps

Why Existing Datasets Fall Short

Synthetic Datasets (LibriMix, WSJMix):

- Artificial mixing of isolated utterances
- Different acoustic conditions per speaker
- Lack of natural room reverberation
- Read speech vs. spontaneous conversation
- Missing reactive overlaps and turn-taking

Real-World Challenges:

- Shared acoustic environment
- Natural loudness relationships
- Ambient noise and reverberation
- Spontaneous speech dynamics
- Complex speaker interactions

Our Contribution: REAL-T Dataset

First Conversation-Centric Real TSE Dataset

- **Multi-lingual**: Mandarin and English data
- **Multi-genre**: Diverse scenarios and styles
- **Multi-enrollment**: Multiple enrollment utterances per speaker
- **Real-world**: Natural conversational dynamics

Key Features:

- Extracted from speaker diarization datasets

Real-World Challenges:

- Infrequent speaker overlaps
- Non-continuous speaking patterns
- Short-term conversational segments
- Environmental noise and non-speech vocalizations

Corpus Selection Strategy

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

Table: Selected datasets for REAL-T construction

Selection Criteria

- Public availability and licensing
- High-quality ASR transcriptions
- Diverse overlap percentages (4.95% - 33.92%)
- Multi-lingual coverage (Mandarin/English)
- Different acoustic scenarios (meetings, dinner parties)

Data Preprocessing Pipeline

Audio Processing:

- Single-channel extraction for microphone arrays
- Channel averaging for distributed arrays
- Far-field recording preference
- Consistent acoustic representation

Transcription Processing:

- Whisper-large-v2 normalization
- Removal of noise tags and punctuation
- RTTM-like format with transcript enrichment
- Standardized evaluation metrics

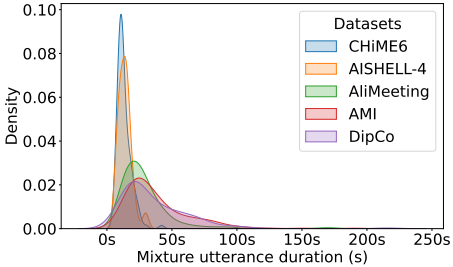


Figure: Distribution of mixture utterance durations across datasets

TSE Data Construction Process

Mixture Utterances:

1. Sort segments by start time
2. Detect overlapping segments
3. Merge overlapping boundaries
4. Calculate overlap statistics
5. Filter by minimum 5s overlap

Enrollment Utterances:

- Extract from non-overlapping segments
- Minimum 5-second duration
- Up to 5 utterances per speaker
- Random selection for balance

Quality Control

- Semantic completeness preservation
- Duration outlier exclusion ($>$100s)
- Speaker proportion validation
- Transcript content verification

Experimental Setup

TSE System:

- BSRNN-based model from WeSep
- Trained on VoxCeleb1 dataset
- Open-source implementation
- Speaker encoder: ECAPA-TDNN

ASR Systems:

- English: Whisper-large-v2
- Mandarin: FireRedASR-AED-L
- Evaluation metric: Token Error Rate (TER)
- Normalized transcriptions

Evaluation Strategy

Since ground truth clean sources are unavailable, we evaluate TSE performance through ASR accuracy on extracted speech

BASE Test Set Analysis

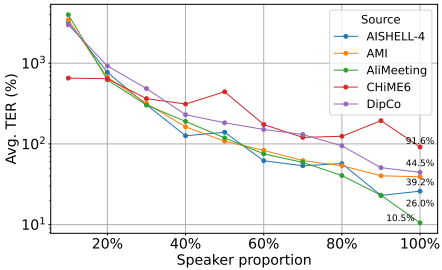


Figure: TER vs. Speaker Proportion Analysis

Key Findings

- Extremely poor performance below 20% threshold
- Some speakers produce non-speech vocalizations
- Need for meaningful speech content filtering

BASE vs. PRIMARY Test Set Performance

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

Table: Average TER (%) on BASE and PRIMARY test sets

Performance Gap

- BASE set shows extremely poor performance (109-185% TER)
- PRIMARY set provides more manageable evaluation (40-92% TER)
- Significant improvement through careful data filtering
- Real-world complexity remains challenging

PRIMARY Test Set Statistics

Category	Source	Lang	T. Dur (min)	Ovl Dur (min)	# Utt	Avg. Ovl. Ratio	# Test
By source							
	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
By language							
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: PRIMARY test set statistics

Dataset Characteristics

- Total: 156.03 minutes, 528 utterances, 1991 test samples
- Balanced language distribution (English: 60%, Chinese: 40%)
- Average overlap ratio: 0.49 (49% overlap)
- Controlled difficulty for meaningful evaluation

Distribution Analysis

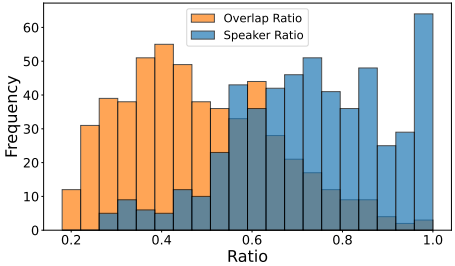


Figure: Distribution of overlap ratio and speaker proportion

Key Observations

- Wide range of overlap ratios (0.2 - 0.8)
- Speaker proportions vary significantly
- Balanced distribution across different scenarios
- Representative of real-world conversational complexity

Impact of Speaker Count on Performance

Spk #	Lang	Avg. TER (%)	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: Average TER (%) and duration by language and speaker count

Performance Trends

- Performance degrades with increasing speaker count
- 2-speaker scenarios show best performance but limited data
- 3-4 speaker mixtures dominate the dataset

Enrollment Utterance Impact

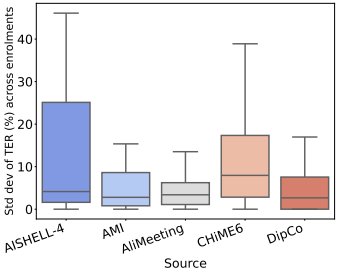


Figure: Distribution of TER standard deviation across different enrollments

Key Findings

- Enrollment utterance choice significantly influences performance
- Large variance observed in AISHELL-4, CHiME6, DipCo datasets
- Speaker embeddings show >50% similarity between good/poor enrollments
- Background noise in enrollment affects extraction quality

Model Performance Comparison

Table: Comparison of model performance on Libri2Mix and PRIMARY test set

Model	Training data	Libri2Mix	PRIMARY Test set	
		SI-SDR(dB)	zh(%)	en(%)
TSELM-L	Libri2Mix-360	/	331.73	192.39
USEF-TFGridnet	Libri2Mix-100	**18.05**	67.98	87.27
BSRNN	Libri2Mix-100	12.95	81.74	91.20
	Libri2Mix-360	16.57	69.80	73.61
	VoxCeleb1	16.50	**57.61**	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	**66.46**

Performance Analysis & Insights

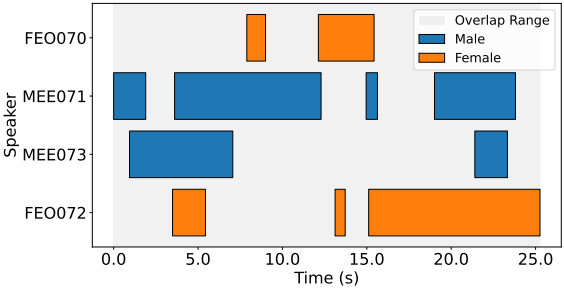
Key Observations

- **TSELM-L:** Generative approach shows poor results, especially on Chinese
- **Language Dependency:** Strong language bias in generative methods
- **Training Data Impact:** VoxCeleb1-trained models generalize better to real-world data
- **Model Comparison:** BSRNN_HR generally outperforms BSRNN

Dataset Difficulty

- CHiME6 and DipCo show worst performance
- Dinner party scenarios more challenging than meetings
- More noise, complex environments, multi-room conditions
- USEF-TFGridnet overfits to Libri2Mix-100

Challenging Example Analysis



Case Study: EN2002a_mixture_0.00_25.26

- **FEO070:** 311.11% TER (lowest speaker ratio)
- **MEE071:** 201.19% TER (primary speaker, extracted as MEE073)
- **MEE073:** 31.58% TER (best performance, higher initial proportion)
- **FEO072:** 89.70% TER (concentrated toward end)

Summary

Contributions

1. **REAL-T Dataset:** First conversational-centric TSE dataset from real diarization data
2. **Comprehensive Analysis:** Identified real-world challenges in TSE
3. **Benchmark Evaluation:** Revealed significant performance gaps between synthetic and real data
4. **Open Source:** All datasets, benchmarks, and metadata will be publicly available

Key Findings

- Existing TSE models show significant performance degradation on real-world data
- Simulated datasets fail to capture real conversational complexity
- Enrollment utterance quality significantly impacts extraction performance
- Real-world scenarios pose unique challenges not addressed by current approaches

Future Work

Dataset Expansion

- More diverse conversational scenarios
- Additional languages and acoustic environments
- Enhanced metadata and annotations

Model Development

- Robust TSE models for real-world conditions
- Better handling of enrollment variability
- Improved performance on challenging scenarios

Thank You!

Questions & Discussion

REAL-T Dataset: `https://real-tse.github.io`

Contact: shuaiwang@nju.edu.cn