# The IACAS-Thinkit System for CHiME-7 Challenge

*Lingxuan Ye*[1,2], *Haitian Lu*[1,2], *Gaofeng Cheng*[1,2], *Yifan Chen*[1,2], *Zengqiang Shang*[1,2], *Xuyuan Li*[1,2]

[1]Key Laboratory of Speech Acoustics & Content Understanding, Institute of Acoustics, CAS, China
[2]University of Chinese Academy of Sciences, Beijing, China

{yelingxuan, luhaitian, chenggaofeng, chenyifan, shangzengqiang, lixuyuan}@hccl.ioa.ac.cn

## Abstract

This paper reports the IACAS-Thinkit's system for the 7th CHiME challenge's task 1: distant automatic speech transcription and segmentation with multiple recording devices. Our system includes training data augmentation, target speaker voice activity detection (TS-VAD) based speaker diarization (SD), time-domain speakerbeam based single channel target speaker extraction (TSE), guided source separation (GSS) based multi-channel speech separation and WavLM based speech recognition. Evaluated on the CHiME-7 evaluation set, our system for the main track achieves 25.0% macro-average Diarization-attributed Word Error Rate (DA-WER), with an absolute reduction of 30.27% over the baseline system; our system for the far-field acoustic robustness sub-track achieves 20.5% macro-average DA-WER, with an absolute reduction of 13.75% over the baseline system.

## 1. System Overview

The CHiME-7 Distant Automatic Speech Recognition (DASR) task focuses on achieving precise speech recognition and speaker diarization under challenging far-field multi-device conditions [1]. In order to address this formidable challenge, our system is structured around three primary modules, namely, speaker diarization, speech separation, and automatic speech recognition. The speech separation module can be further dissected into two components: single-channel target speaker extraction (TSE) and multi-channel speech separation. We also enhance the accuracy of our speech recognition outputs by rescoring them with a language model. It is noteworthy that our system is specifically tailored to cater to both the main track and sub-track of the CHiME-7 DASR task. An illustrative representation of our system can be found in Figure 1.

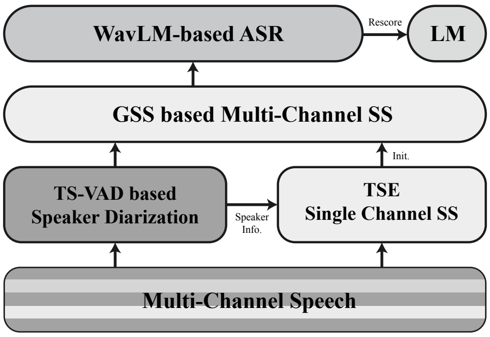

Figure 1: *Overview of the IACAS-Thinkit's system*

## 2. Speaker diarization

### 2.1. Model Configuration

Our TS-VAD differs from the original TS-VAD [2] using i-vector [3] by employing ECAPA-TDNN [4] based x-vector as the speaker embedding. Firstly, we use an ECAPA-TDNN which has the same structure as the speaker embedding model to extract the frame-level speaker embeddings. A statistical pooling layer is employed on the frame-level speaker embeddings every ten frames because speaker diarization does not need such a high temporal resolution [5]. The output of the statistical pooling layer is named segment-level speaker embeddings. Secondly, the segment-level speaker embeddings are concatenated with the target-speaker embeddings. It is worth noting that the target-speaker embeddings are estimated iteratively in the same way as [2]. A two-layer BiLSTM detects the states of each speaker separately then the detection states are concatenated and fed into a BiLSTM layer to find the relationship between different speakers. Finally, a linear layer with sigmoid function outputs per frame target-speaker voice activity. The system diagram is depicted in Figure 2.

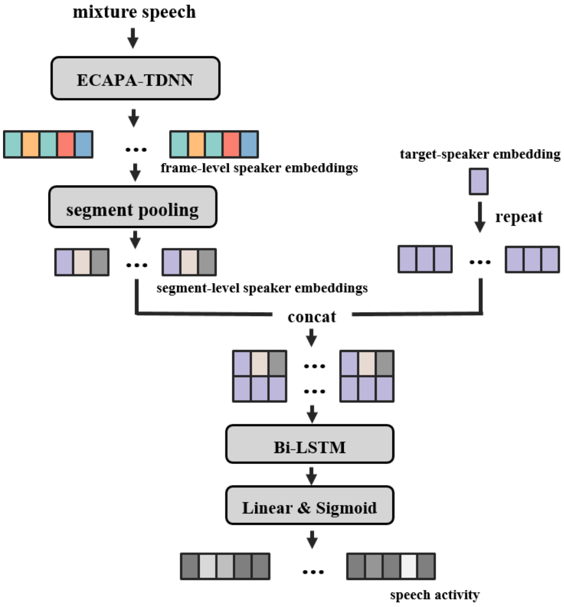

Figure 2: *The system diagram of ECAPA-TDNN based TS-VAD.*

### 2.2. Training Data and Training Details

#### 2.2.1. *Speaker embedding model*

In order to achieve better generalization on the CHiME-7 dataset, we generate a speaker identification training set by extracting single-speaker speech from CHiME-7 training set according to the manual annotation. The speaker embedding model is pretrained on VoxCeleb1 [6] and VoxCeleb2 [7] and finetuned on Voxceleb1 and CHiME-7 training set.

#### 2.2.2. *TS-VAD*

Data augmentation plays an important role in the performance of TS-VAD. On the one hand, we use slot-filling to generate simulated training data. We remove speech regions in the CHiME-6 [8] training set and fill the regions with speech from LibriSpeech [9], VoxCeleb1 [6] and VoxCeleb2 [7]. Additionally, we add background noise and reverberation to the simulated training data. The background noise is extracted from the non-speech regions from the CHiME-6 training dataset, while the room impulse response is sampled from the SLR28 RIR and Noise Database [10]. We generate 6000 hours of simulated training data in this way. On the other hand, we perform random orthogonal transforms to the frame-level speaker embeddings and target-speaker embeddings to avoid overfitting to some specific regions in the speaker embedding space [5][11]. The probability of performing random orthogonal transforms to each batch is set to 0.4.

There are three stages in the training phase [12]. In the first stage, the front-end ECAPA-TDNN's parameters are copied from the pretrained speaker embedding model. We freeze the front end and train the backend BiLSTM layers with a learning rate of 0.0001 for 20 epochs on the simulated training dataset. In the second stage, we unfreeze the front-end model and train the whole model on the simulated training dataset for another 2 epochs with a learning rate of 0.0001. In the third stage, we finetune the whole model on the CHiME-6 training set for 2 epochs with a learning rate of 0.00001.

In addition, we perform self-supervised domain adaptation on the development set and evaluation set. We tune the model for 30 steps with a learning rate of 0.00001 on each session in the development set and evaluation set. The training target is the DOVER-lap output and the target-speaker embedding is also extracted from the diarization output.

### 2.3. Inference

We employ several post-processing strategies in the inference phase of TS-VAD, including merging two speech segments separated by a pause shorter than 0.6 seconds, deleting all speech segments shorter than 0.2 seconds, and binarization by a threshold of 0.4 and 5-tap median filtering. In order to make use of multi-channel audio in the development set and evaluation set, we perform DOVER-lap on the output of different channels [13]. The results of the development set are shown in Table 1. We use the baseline diarization results as the initial diarization. We delete speakers who speak less than 100 seconds and speak ten times less than the speaker who speaks the most in the session. The speaker diarization results are shown in Table 1.

Table 1: *Speaker diarization results on Dev&Eval [1].*

| Methodology | Scenario | Dev | | Eval | |
|---|---|---|---|---|---|
| | | DER | JER | DER | JER |
| **Baseline** | CHiME-6 | 40.0 | 51.1 | 56.3 | 62.5 |
| | DiPCo | 29.8 | 41.4 | 27.9 | 40.9 |
| | Mixer 6 | 16.6 | 22.8 | 9.3 | 11.0 |
| | Macro | 28.8 | 38.5 | 31.2 | 38.2 |
| **Ours** | CHiME-6 | 25.2 | 29.4 | 27.3 | 31.4 |
| | DiPCo | 22.1 | 23.1 | 22.4 | 28.0 |
| | Mixer 6 | 14.7 | 20.8 | 7.3 | 8.0 |
| | Macro | 20.7 | 24.5 | 19.0 | 22.5 |

## 3. Speech Sepatation

### 3.1. Single Channel Target Speaker Extraction

We perform single-channel target speaker extraction (TSE) with time-domain speakerbeam [14]. We choose time-domain speakerbeam because it is a classical TSE method, and it achieves better performance than the traditional frequency-domain speakerbeam.

The training data is crucial for the performance of the time-domain speakerbeam on the CHiME-7 dataset. At first, we only trained the model on Libri3mix, and it led to bad generalization on the CHiME-7 dataset. In order to solve the problem, for each session in the CHiME-7 development and evaluation set, we extract single speaker segments according to the TS-VAD diarization output and combine them randomly to generate the simulated training set in an on-the-fly manner. As for the model training, we first train time-domain speakerbeam on Libri3mix [15] for 18 epochs. Next, we finetune the model on the simulated training set for 30 epochs. Note that we train a TSE model for each session in the development and evaluation sets. We managed to enhance the TSE model's performance on the CHiME-7 dataset significantly in this manner. After obtaining the single speaker signal from the mixed signal with the time-domain speakerbeam, we use it to calculate the time-frequency spectral mask for each speaker.

### 3.2. Multi Channel Source Separation with GSS

For each session on dev/eval separately, based on the RTTM results provided by the SD system, we extract the speech of each speaker without overlap and concatenate them as the registered speech for the TSE model. For each utterance extracted by the SD system, we first use the TSE model to separate each speaker's speech. Then, we get the time-frequency spectral mask for each speaker by performing FFT to initialize the GSS iteration.

We present the results of applying TSE init to the original in Table 2. We show the results on the CHiME-6 Dev set. The ASR model used is trained on the original training sets.

We can see that using TSE's pre-separation result to initialize GSS could bring about 4% relative WER reduction.

## 4. Automatic Speech Recognition

### 4.1. Acoustic Model (AM)

For AM, we adopt the WavLM (Large)[16] for all our experiments since its performance beats the other SSL models. We

Table 2: *Contribution of Speech Separation to WER on CHiME-6 Dev.*

| Seperation | WER% | | |
|---|---|---|---|
| | S02 | S09 | ALL |
| GSS | 38.1 | 33.6 | 35.3 |
| +TSE init | 37.6 | 31.6 | 33.9 |

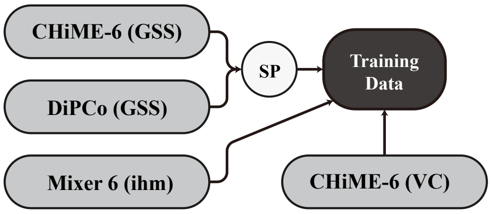

Figure 3: *Training Sets of the AM*

first finetune the pretrained model on the re-separated training sets and then perform continuous unsupervised adaptation[17] for each session on eval sets separately.

The composition of training sets of the AM is shown in Figure 3. We first re-separate the development sets and put half of the original development sets into the training set to train a trial ASR system to determine the best number of steps for training. Then, we add all the speech from the dev sets to train the model to such a number of steps to prevent over-fitting. For chime6 and dipco, we used the data enhanced by GSS and performed 3-times speed perturbation on these sets (29.6h * 3 from the original chime6 training set and 6.1h * 3 from the original chime6 dev set). For Mixer6, we choose the mdm and ihm speech from the original training set without augmentation and subsample to 54.8h, plus the enhanced dev set without speed perturbation (10.9h). For DiPCO, we use all the dev sets with 3-times speed perturbation, and 2-times upsample (3.3h * 3 * 2). To extend the robustness of the ASR system across speakers, we also added 20h generated data from voice conversion (VC). The VC model is trained on the Librispeech corpus and converts utterances in CHiME6's ihm training set to speakers in the Librispeech corpus. For training the trial ASR system, we use sessions of S09, S33, S29, 20090714-134807-LDC-120290, 20090716-155120-LDC-120269, 20090717-113617-LDC-120278, 20090717-133033-LDC-120311, 20090722-115429-LDC-120271, 20090722-154451-LDC-120225, 20090723-111806-LDC-120290, 20090729-155715-LDC-120311, 20090803-111429-LDC-120225, 20090803-120934-LDC-120271, 20090804-165853-LDC-120269 and 20090805-110532-LDC-120225 as the Resep-Dev set and the remaining as the training set. We figure out the best number of training steps of the AM on our Resep-Dev set, and finally, we add all of the 213.5h speech both from training sets and development sets for training and trained such number of steps (40k updates) to circumvent over-fitting. The number of adaptation steps is also determined in such a way, and we submitted the results of adapted 30, 50, and 90 steps as the three system results.

### 4.2. Language Model (LM)

For LM, we use all the text from the training set and development sets (1.2M words) to train a 4-gram model. The 4-gram model is interpolated with a 4-gram model trained on Librispeech's text (9.6M words). We also trained a cross-utterance transformer LM on the training text from original splits (1.0M words). During training and inference of the transformer LM, We concatenate the past two segments of the same speaker's text and add an additional symbol between utterances to leverage the long-context information. The final decoding is performed by first beam-search on CTC posteriors combined with the n-gram model. We decode 60 best results and finally rescore the results to get 1-best with transformer LM. The results are shown in Table 3. The ASR model used is trained on the original training sets.

Table 3: *Contribution of the long-context NNLM to WER on CHiME-7 Dev.*

| Seperation | WER% | | | |
|---|---|---|---|---|
| | CHiME-6 | DipCo | Mixer 6 | Macro |
| WavLM+ngram | 35.7 | 39.4 | 14.3 | 29.8 |
| +NNLM | 35.3 | 38.6 | 14.4 | 29.4 |

## 5. Results & Discussion

The ASR results of the first system (used to determine the best number of training and adaptation) on our re-separated dev set (Resep-Dev) are shown in Table 4. The final system is trained on all the training and development sets.

Table 4: *The trial ASR system's results on the re-separated Dev. set*

| Scenario | Resep-Dev (WER%) | |
|---|---|---|
| | Sub-Track | Main-Track |
| CHiME-6 | 18.0 | 35.7 |
| DiPCo | 25.8 | 32.3 |
| Mixer 6 | 10.7 | 12.5 |
| Macro | 18.2 | 26.8 |

We demonstrate the final results of our system for the main track and the far-field acoustic robustness sub-track in Table 5.

Table 5: *System performance results on the evaluation sets of CHiME-7*

| Scenario | WER% | | | |
|---|---|---|---|---|
| | Main Track | | Sub-Track | |
| | Baseline | Ours | Baseline | Ours |
| CHiME-6 | 77.4 | 31.1 | 35.5 | 23.9 |
| DiPCo | 54.7 | 25.4 | 36.3 | 20.5 |
| Mixer 6 | 33.7 | 18.5 | 28.6 | 17.0 |
| Macro | 55.3 | 25.0 | 33.4 | 20.5 |

Overall, our system achieved a relative WER reduction of 54.8% on the main track and a relative WER reduction of 38.6% on the far-field acoustic robustness sub-track. By comparing the results from the main track and the sub-track, we could find SD benefits CHiME-6 and DiPCo sets most but have limited impact on the Mixer 6 set since the Mixer 6 set has a lower overlap ratio. The CHiME-6 set is the most challenging set among the three sets, both for the diarization system and the ASR system.

# 6. Conclusions

This paper outlines the IACAS-Thinkit system for the CHiME-7 challenge's task 1: distant automatic speech transcription and segmentation with multiple recording devices. Our system encompasses three integral components: speaker diarization, speech separation, and automatic speech recognition. A key observation from our efforts is the paramount importance of data augmentation within the context of this challenge. Furthermore, we leverage self-supervised domain adaptation techniques, substantially enhancing our final results. Additionally, we introduce certain modifications to the model architecture, which also leads to noticeable improvements. In summary, our system's performance in the CHiME-7 DASR task was notable. In the main track, our system secured the second position with a DA-WER of 26.8% on the development set and 25.0% on the evaluation set. In the far-field acoustic robustness sub-track, our system achieved a DA-WER of 18.2% on the development set and 20.5% on the evaluation set, positioning us at the fourth rank.

# 7. References

[1] S. Cornell, M. Wiesner, S. Watanabe, D. Raj, X. Chang, P. Garcia, Y. Masuyama, Z.-Q. Wang, S. Squartini, and S. Khudanpur, "The chime-7 dasr challenge: Distant meeting transcription with multiple devices in diverse scenarios," *arXiv preprint arXiv:2306.13734*, 2023.

[2] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," in *Proc. Interspeech 2020*, 2020, pp. 274–278.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[5] W. Wang, Q. Lin, D. Cai, and M. Li, "Similarity measurement of segment-level speaker embeddings in speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2645–2658, 2022.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[7] 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.

[8] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[10] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[11] Q. Li, F. L. Kreyssig, C. Zhang, and P. C. Woodland, "Discriminative neural clustering for speaker diarisation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 574–581.

[12] W. Wang, D. Cai, Q. Lin, L. Yang, J. Wang, J. Wang, and M. Li, "The dku-dukeece-lenovo system for the diarization task of the 2021 voxceleb speaker recognition challenge," *arXiv preprint arXiv:2109.02002*, 2021.

[13] D. Raj, L. P. Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "Dover-lap: A method for combining overlap-aware diarization outputs," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 881–888.

[14] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 691–695.

[15] 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.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, p. 1505–1518, Oct 2022, arXiv:2110.13900 [cs, eess].

[17] Y. Higuchi, N. Moritz, J. L. Roux, and T. Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," *arXiv preprint arXiv:2106.08922*, 2021.