The NAIST System for the CHiME-8 NOTSOFAR-1 Task

Yuta Hirano, Mau Nguyen, Kakeru Azuma, Jan Meyer Saragih, Sakriani Sakti

Introduction

Previous challenges focused sorely on high accuracy

Faster Inference Speed

Higher Acuraccy

➡ **We aim to reduce inference speed while improving recognition accuracy.**

Inference time of the baseline system mostly comes from the ASR module

➡ **We mainly worked on the ASR module.**

Proposed System

The baseline system

Conformer CSS → Whisper large → Nemo Diarizer

Proposed system* * Duplicated modules are grayed out

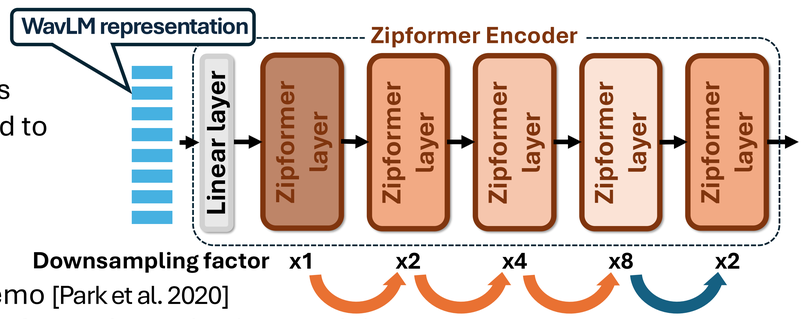
WPE (Optional) → Conformer CSS → WavLM large → Zipformer-T → RNN LM (Optional) → Nemo Diarizer

Continuous Speech Separation

- Conformer CSS [Chen, et al. 2020]
  - Checkpoints from the baseline system was used
- **Block-wise WPE dereverberation**
  - Applied to 3.0 sec of audio chunks
  - Implementation in nara_wpe was used

Automatic Speech Recognition

- **WavLM-large feature extractor [Chen, et al. 2021]**
  - The output of the last (24th) layer was used
- **Zipformer transducer [Zengwei, et al. 2023]**
  - Fast and memory-efficient because of U-Net-like downsampling and upsampling operations
  - Better performance compared to other transformers
  - Word timestamps
    - Obtained by referring to input audio frame indexes
    - Word duration was limited to 1.0 sec
- **RNN LM**
  - Performs shallow fusion during decoding



Speaker Diarization

- Word-nmesc dierizer from Nemo [Park et al. 2020]
  1. Extract speaker embedding for each word using word timestamps
  2. Assign speaker labels by performing spectral clustering

Table 1: *WER (%)on AMI corpus*

Model	Parameter size	IHM	SDM
Zipformer-T (Fbank)	70.4 M	17.6	32.8
Zipformer-T (WavLM-large 24th)	70.0 M	**13.2**	**26.3**
Whisper large	1.6 B	16.9	36.4

Experiment

1. Dataset

- Two datasets for model training
- Overall, we tried to **increase the ratio of meeting recordings**

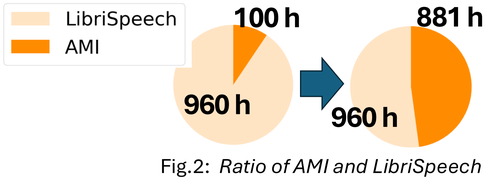
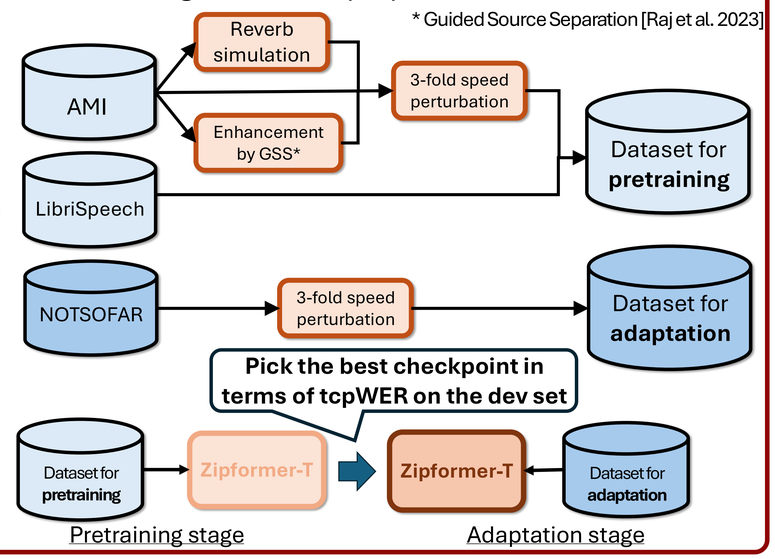


Fig.2: *Ratio of AMI and LibriSpeech*

2. Training

- Two NVIDIA A100 40GB GPUs
- Pruned RNN-T loss [Kuang, et al. 2022]
- SpecAugment

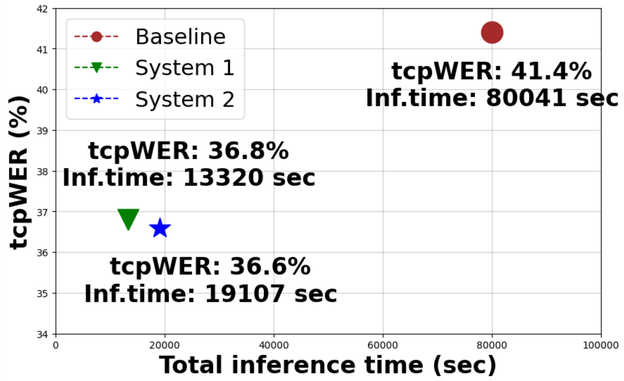
Fig.1: *Dataset preparation flow*



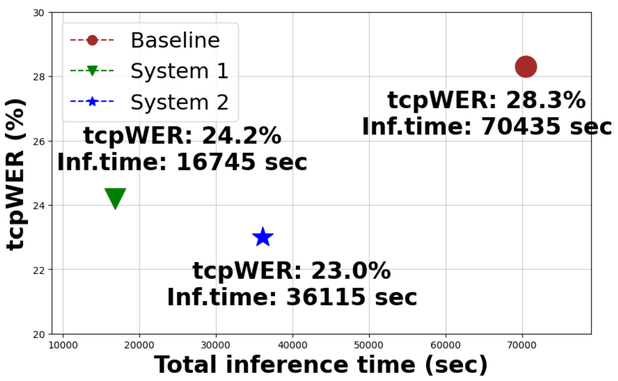
Result

Component	Single-channel track		Multi-channel track	
	System 1	System 2	System 1	System 2
WPE				✓
Conformer CSS	✓	✓	✓	✓
WavLM-large	✓	✓	✓	✓
Zipformer transducer	✓	✓	✓	✓
RNN LM		✓		
Nemo word nmesc diarizer	✓	✓	✓	✓

Single-Channel Track



Know-Geometry Multi-Channel Track



- We reduced **tcpWER** by 11.1 - **11.6** % (🥉 **3rd place** out of 6 teams)
- We reduced **inference time** by 76.1 – **83.4** %
- Shallow fusion by RNNLM resulted in only 0.2 % tcpWER reduction

Brought by our ASR module only

- We reduced **tcpWER** by 14.5 - **18.7** % (**4th place** out of 10 teams)
- We reduced **inference time** by 48.7 – **76.2** %
- WPE dereverberation reduced tcpWER by 1.2 %, but increased the inference time by nearly 20000 sec.

Human-AI Interaction Laboratory, Nara Institute of Science and Technology