ICASSP2024 ICMC-ASR Workshop

The USTC-NERCSLIP Systems for the ICMC-ASR Challenge

Presenter: **Yichi Wang**

NERC-SLIP, University of Science and Technology of China (USTC)

Anhui iFLYTEK Universal Language Technology, China

Huawei Technology, China

04/17/2024

The USTC-NERCSLIP Team

Minghui Wu
(USTC)

Jie Zhang
(USTC)

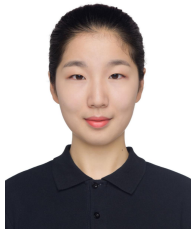
Yanyong Zhang
(USTC)

Xin Fang
(iFlytek)

Haitao Tang
(iFlytek)

Chongliang Wu
(iFlytek)

Yongchao Li
(iFlytek)



Luzhen Xu
(USTC)

Yichi Wang
(USTC)

Zhengzhe Zhang
(USTC)

Tongle Ma
(iFlytek)

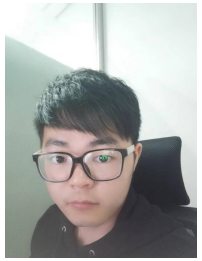
Yanyan Yue
(iFlytek)

Ruizhi Liao
(iFlytek)

Jiachen Liu
(iFlytek)

Haoyin Yan
(USTC)

Jintao Zhao
(iFlytek)



Hongliang Yu
(iFlytek)

Yue Zhang
(Huawei)

The ICASSP2024 In-Car Multi-Channel Automatic Speech Recognition Challenge (ICMC-ASR)

Track I: Automatic Speech Recognition (ASR)

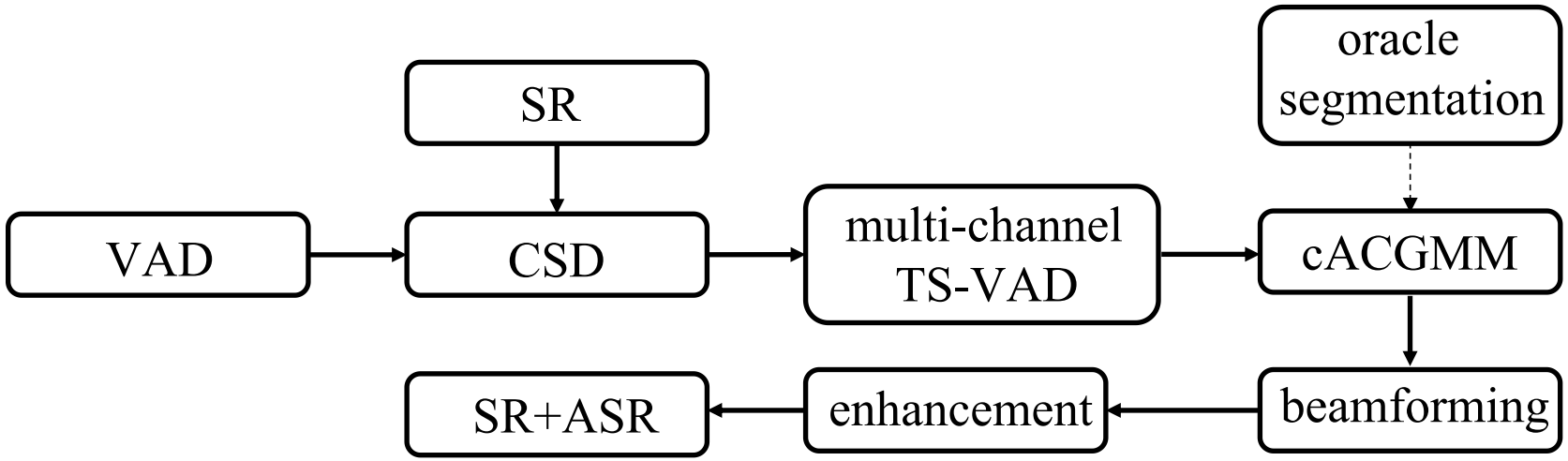
Participants will receive the oracle segmentation of the evaluation set. The goal is to develop ASR systems based on the multi-channel multi-speaker speech data. Participants need to devise algorithms that can effectively fuse information across different channels, suppress background noise, and handle multi-speaker overlaps.

Track II: Automatic Speech Diarization and Recognition (ASDR)

Participants will not receive any prior or oracle information during evaluation (e.g., segmentation and speaker label for each utterance, total number of speakers in each session). Participants in this track are required to design automatic systems for both speaker diarization (identifying who is speaking when) and transcription (converting speech to text).

Both pipeline and end-to-end systems are acceptable, allowing for flexibility in system design and implementation.

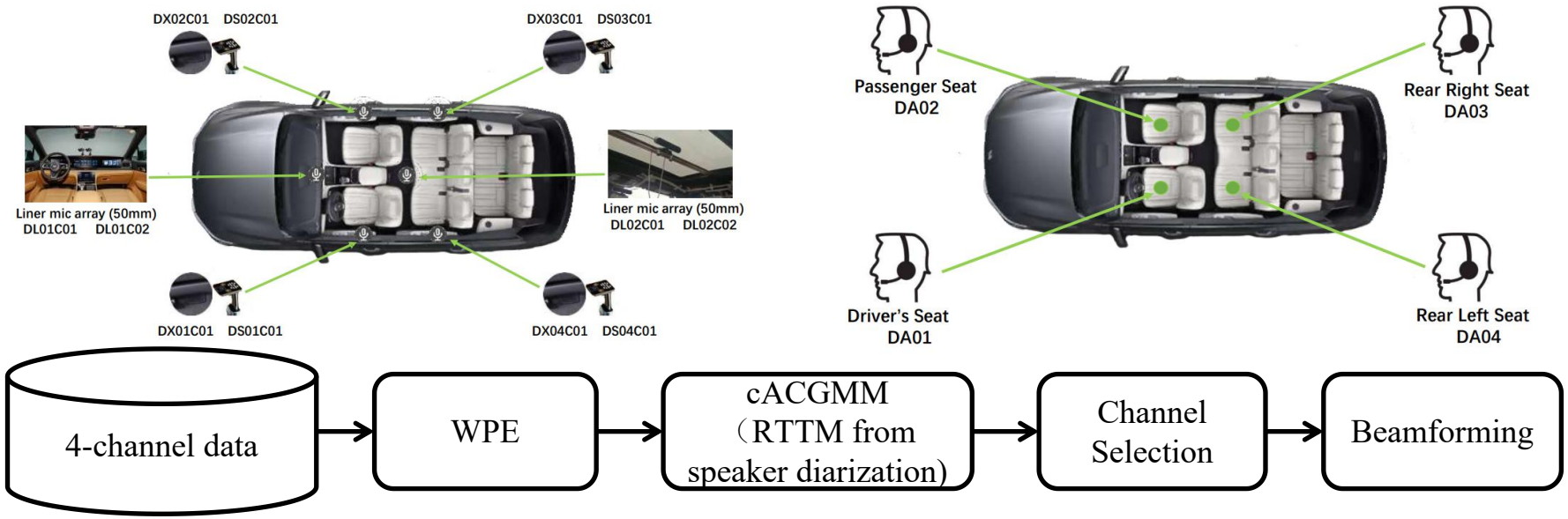
Overall Framework



- **VAD (Voice Activity Detection):** removes silent segments to obtain "spoken when."
- **SR(Speaker Recognition):** providing speaker information.
- **CSD (Cluster-based Speaker Diarization):** performs unsupervised classification to obtain "who spoken when."
- **TS-VAD (Target Speaker Voice Activity Detection):** extracts i-vectors for CSD, combined with far-field audio, to achieve better "who spoken when."
- **cACGMM(complex Angular Central Gaussian Mixture Model):** cACGMM can model not only rotationally symmetrical but also elliptical distributions.

Front-end Processing

Channel Selection + Guided Source Separation (GSS) [1]



- **Channel selection based multi-source sound localization** using energy and phase differences: selecting the channel for target speaker by using a multi-source sound localization module.

- **The average iterative algorithm** using in MVDR beamformer

[1] Ruoyu Wang, Maokui He, et al., “The ustc-nercslip systems for chime-7 challenge,” in Proc. CHiME 2023, 2023, pp. 13–18.

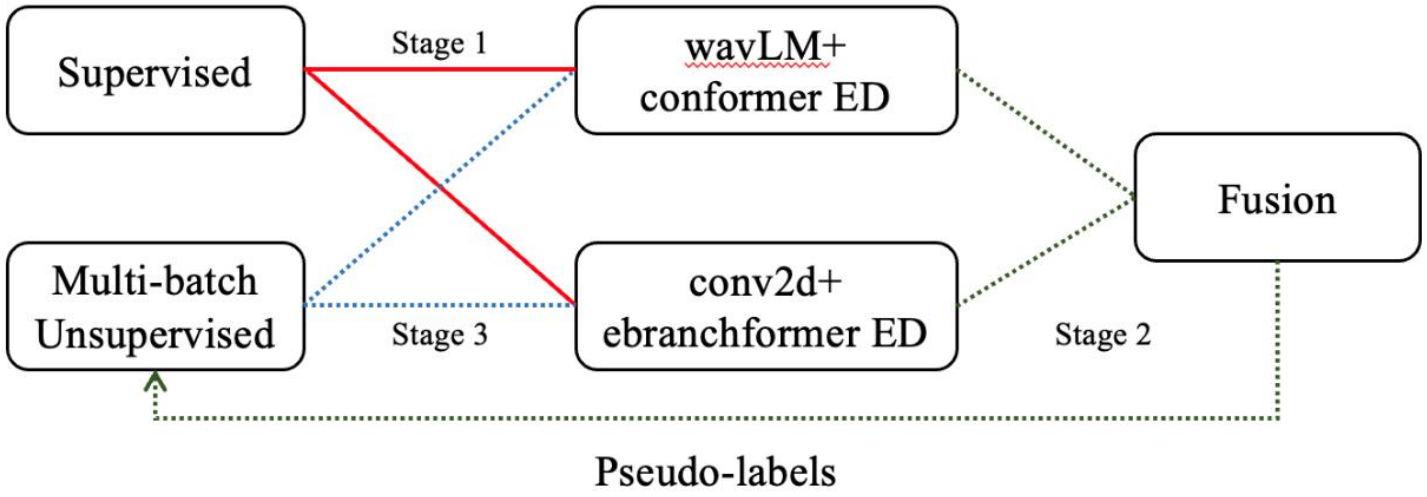
Data Description

Dataset	Data type	Label type	Hours (h)
ICMC	Far+Near+speed×3	supervised	386
ICMC+addNoise	Far+speed×3		1544
3D-SPEAKER	Near	unsupervised (text is not allowed to use)	1124
AliMeeting	Far+Near		236
AISHELL-4	Far+Near		240
Aidatatang	Near		200
MagicData	Near		180
KeSpeech	Near		1542
WenetSpeech	Near (drama,talk,interview)		5500

- The official ICMC-ASR labelled data
- The external openSLR unlabelled data following the official rules [2];

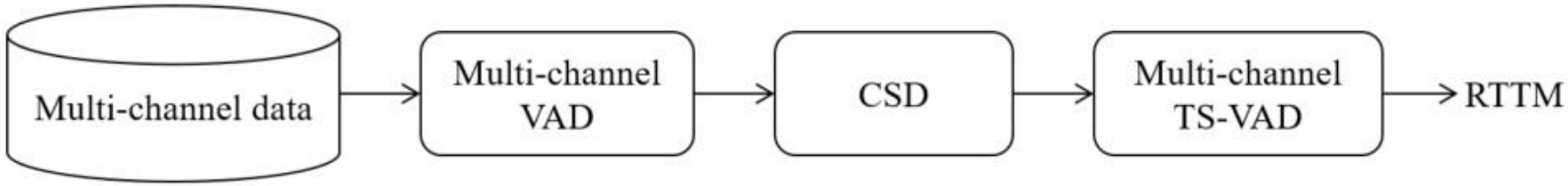
[2] Kazuya Kawakami, Luyu Wang, et al., “Learning robust and multilingual speech representations,” arXiv preprint arXiv:2001.11128, 2020.

Pseudo-label Generation for Unsupervised Data



- **Unsupervised data:** we propose an iterative pseudo-label generation (PLG) method;
- **Self-supervised learning representation (SSLR):** extracted by the adaptive wavLM model [2];
- **Pseudo-label generation:**
  - A small amount of supervised data B_0 is initially used to train and fuse into PLG_0;
  - The unsupervised data is then split into different batches, say B_1, ..., B_N;
  - The pseudo-labels for the unsupervised data in B_N are generated by PLG_{N-1} in a cyclic iterative way;
  - PLG_N is trained and fused by B_N .

Multi-stage speaker diarization system

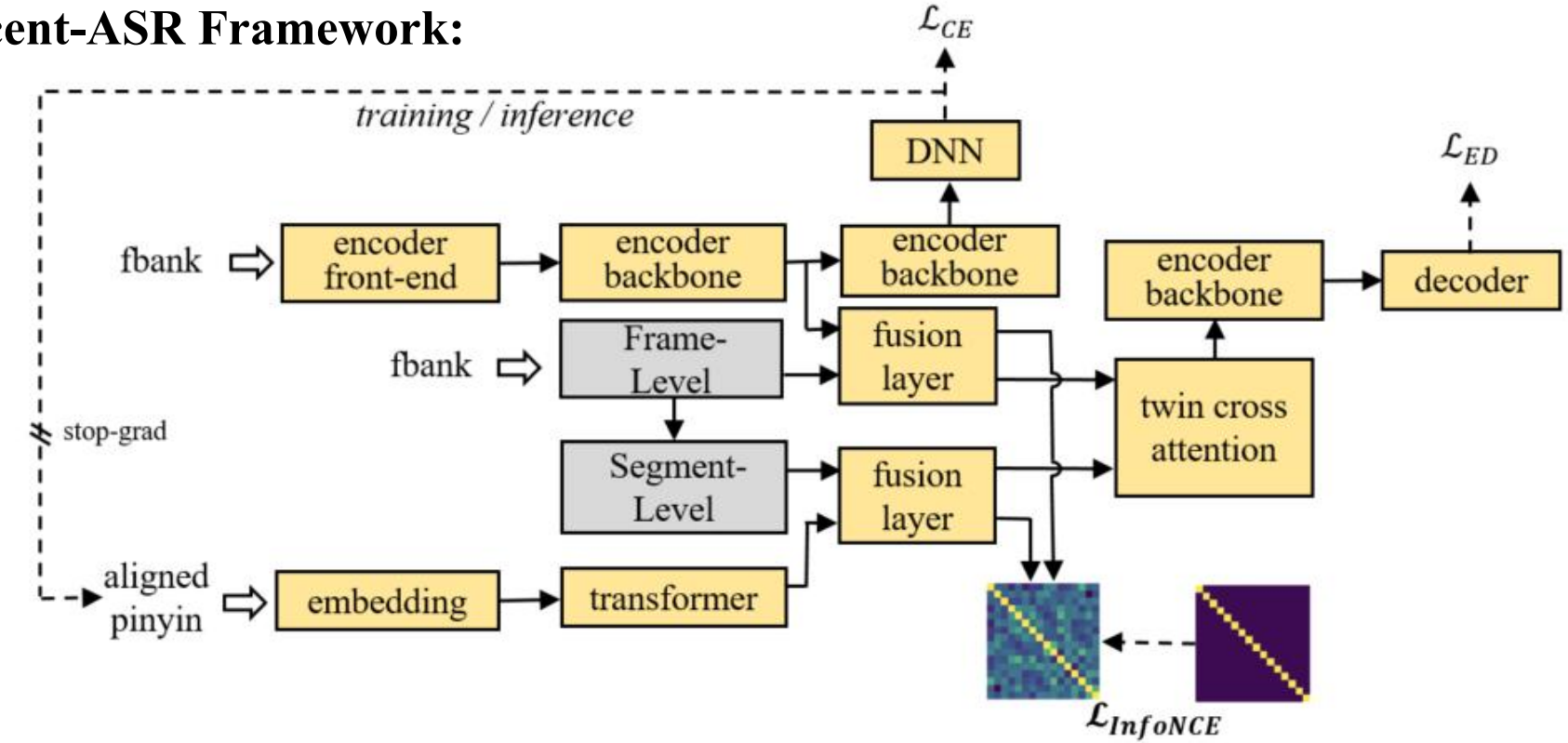


- **Multi-channel voice activity detection (VAD)**
  - Accurately identify the presence of speech across multiple audio channels.

- **Clustering-based speaker diarization (CSD)**
  - Generate the initial clustering results of the multi-channel data.

- **Multi-channel target-speaker VAD (MC-TS-VAD)**
  - The fusion of the SSLR-based x-vector and the i-vector in the MC-TS-VAD module to improve the generalization of speaker embeddings.

Accent-ASR Framework:



- To address the crosstalk that exists between accented Mandarin and standard Mandarin.
- Fine-grained units to capture pronunciation-related accent characteristics.
- Coarse-grained units to learn linguistic information.

Speech Recognition

Accent-ASR :

- **Multi-task learning for aligning Pinyin sequences :**
  - Combines the aligned Pinyin features with the acoustic features of the encoder through twin cross-attention fusion and contrastive learning.
  - The aligned Pinyin features are derived from embedding and transformer modules.

- **Fusion layer :**
  - The frame-segment level speaker information is further introduced to help distinguish coarse-grained units that are generated by speakers of different accents.

- **Accent-ASR fusion model :**
  - Modify and combine the front and back sub-modules of encoder.
  - The front-end module includes conv2d, VGG and gatecnn.
  - The backbone module consists of conformer and ebranchformer [3].

[3] Weitai Zhang, Zhongyi Ye, et al., “The ustc-nelslip offline speech translation systems for iwslt 2022,” in Proc. IWSLT, 2022, pp. 198–207.

Speech Recognition Results (Track 1)

The CER(%) performance of the pseudo-label iterative generation based on the PLG fusion model

Iteration	Data(hrs)	wavLM+ conformer ED	conv2d+ ebranchformer ED	Fusion
B0	396h	29.77	30.64	27.8
B1	3910h	25.32	26.4	23.77
B2	5452h	22.94	21.9	22.87
B3	10952h	22.94(5452h)	21.14	20.83

- As the number of iterations increases, the performance of PLG significantly improves.
- The wavLM-based ED model has competitive results given a relatively small amount of data.
- As the data amount increases to 5000 hours, the wavLM-based ED model becomes inferior to that of the Fbank-based ED model.

Speech Recognition Results (Track 1)

The CER(%) results of ASR single and fusion models on the Track 1 dev and eval sets

ID	Model based on Accent-ASR	Dev	Eval
M0	Official Baseline	32.92	26.24
M1	conv2d+conformer	20.87	-
M2	conv2d+ebranchformer	20.68	-
M3	VGG+ebranchformer	20.31	-
M4	gatecnn+conformer	19.83	-
M5	gatecnn+ebranchformer	18.53	14.72
M6	fusion models based on M1-M5	15.54	13.16

- M1-M5 are single model composed of different encoder front-end and back-end sub-modules.
- M6 denotes the post fusion based on weight adaptation (achieves a relative improvement of 52.8% and 49.8% compared to M0 on the dev and eval sets, respectively).

Track I ASR | Track II ASDR

Rank	Entries	TeamID	CER (%)	Upload Date
1	8	T005	13.16	12-20 18:59
2	24	T078	14.63	12-20 19:33
3	8	T052	14.72	12-20 16:43
4	31	T018	15.46	12-18 21:50
5	5	T024	15.62	12-20 19:31

T005 USTC-iFlyTech

Overall Results (Track 2)

The automatic speaker diarization and recognition (ASDR) results on Track 2 dev and eval sets

Model	Metric	Dev	Eval
Speaker Diarization	DER (%)	10.21	-
M0 Offical Baseline	cpCER (%)	65.90	72.88
M6 Fusion models		16.31	21.48

- In comparison to the official baseline, our system achieves a cpCER of 16.31% on the dev set, which relatively improves by 75.3%.
- On the eval set, our system outperforms the official baseline with a relative improvement of 70.5%.

Track I ASR | Track II ASDR

Rank	Entries	TeamID	cpCER (%)	Upload Date
1	15	T005	21.48	12-20 19:00
2	5	T054	25.88	12-20 16:43
3	7	T001	26.37	12-20 19:59
4	11	T024	26.48	12-20 19:31
5	15	T078	29.33	12-20 19:08

T005 USTC-iFlyTech

Conclusion

- **Channel selection based multi-source sound localization and MVDR beamformer** are important to provide high-quality speech for the downstream tasks of our system, such as speaker diaraization and speech recognition.
- **The fusion of the SSLR-based x-vector and the i-vector in the MC-TS-VAD module** enables the diarization system to fully exploit speaker information and is thus helpful for speaker separation.
- **The post fusion of Accent-ASR model based on weight adaptation** helps improve the E2E ASR performance (a relative improvement of 52.8% compared to official baseline in CER).
- The proposed system achieves the best performance on both tracks of this challenge.

Thanks to the organizers to arrange this grand challenge!

Thanks for listening!

Q&A

{mhwu, httang}@iflytek.com

Please wait the workshop paper for experimental details!