# Cross-speaker encoding network for multi-talker speech recognition

Jiawen Kang, Lingwei Meng, Mingyu Cui, Haohan Guo, Xixin Wu, Xunying Liu, Helen Meng

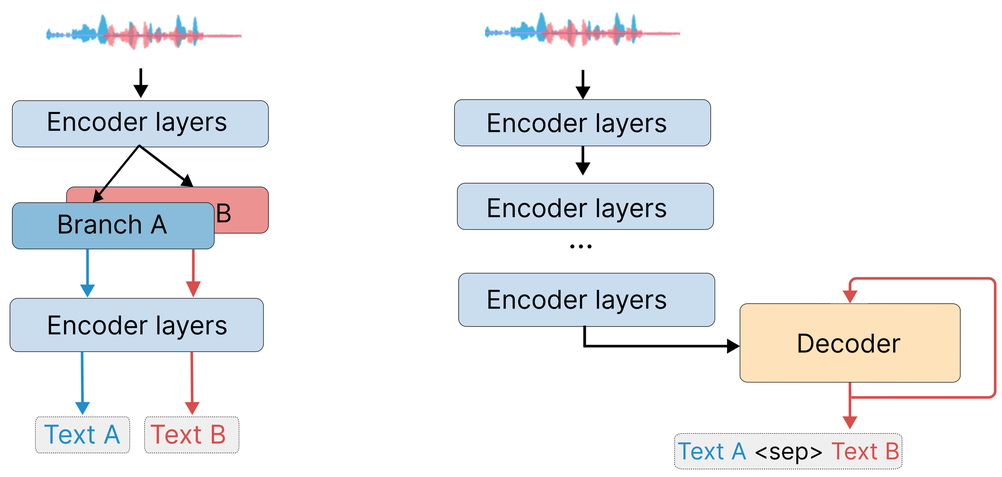
The Chinese University of Hong Kong

## Highlights

- Proposed Cross-Speaker Encoding Network (CSE-Net) for multi-talker speech recognition
- CSE-Net integrated with Serialized Output Training (SOT), combining strengths of Single-input multiple-output and Single-input single-output models models as CSE-SOT-Net
- Achieved state-of-the-art performance on LibrispeechMix across varying overlap rates

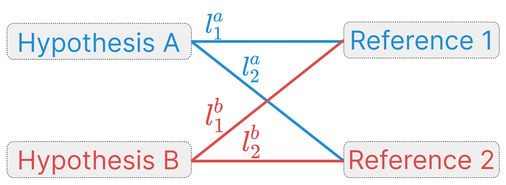
## Introduction

- Architectures for multi-talker speech recognition:




Single-input multiple-output (**SIMO**) Single-input single-output (**SISO**)

- **Single-input multiple-output (SIMO)**
  - Pros: Explicitly models speaker separation
  - Cons: Isolate speakers into separate branches
- **Single-input single-output (SISO)**
  - Pros: Can handle varying number of speakers
  - Cons: No explicit speaker separation modeling
- Loss functions for multi-talker speech recognition:

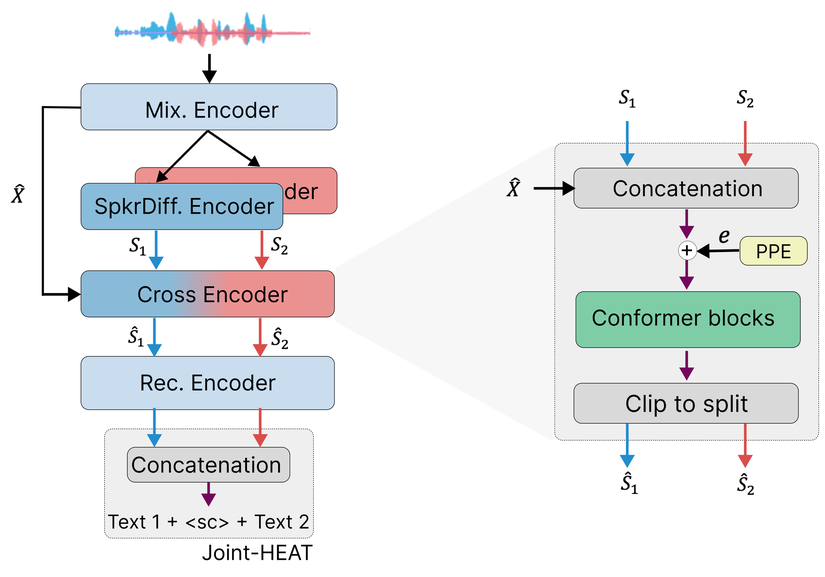



$$L_{PIT} = min(l_1^a + l_2^b,\ l_2^a + l_1^b)$$

$$L_{HEAT} = l_1^a + l_2^b$$

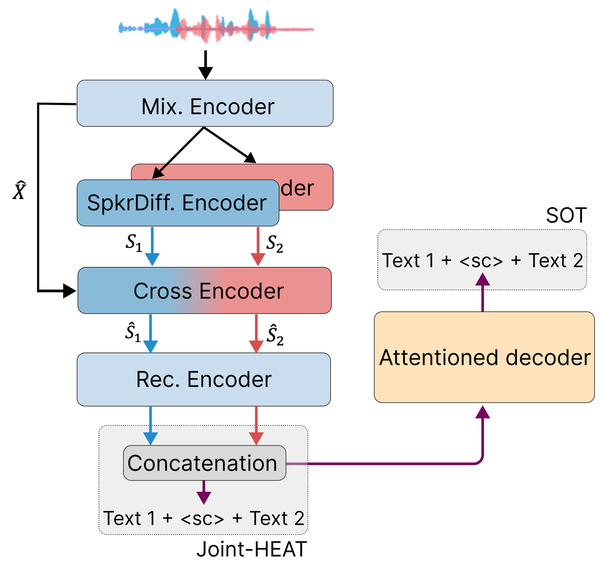
- **Permutation Invariant Training (PIT)**
  - Pros: No prior assumptions on speaker orders
  - Cons: O(n^2) complexity w.r.t. number of speakers
- **Heuristic Error Assignment Training (HEAT)**
  - Pros: O(n) complexity w.r.t. number of speakers
  - Cons: Require additional criterion to determine speaker orders

## Cross-Speaker Encoding network




**Cross-Speaker Encoding Network (CSE-Net)** **Cross Encoder**

- *Cross-encoder* model inter-speaker dependencies and enabling separate branches *condition on each other*
- Cross-encoder concatenated separate representations with a *partition-wise positional embedding (PPE)* indicate positions
- Self-attention in conformer block provides global view cross representations
- Joint-HEAT calculates HEAT loss on concatenated representations to unify branches




**CSE-SOT-Net**

- As Joint-HEAT merged branches, a decoder was cascaded in a joint ctc/attention paradigm
- Serialized output training (SOT) was adopted to transcribe speech one speaker after another
- CSE-SOT-Net integrates SIMO and SISO approaches, leveraging their advantages while mitigating drawbacks

## Experimental setup

- **LibrispeechMix Dataset:**
  - Simulating overlapped speech from LibriSpeech corpus "clean" subset
  - Training set: 1-speaker + 2-speaker speech, with 1.7k hours, 400k utterances
  - Test set: standard LibrispeechMix {1mix, 2mix, 3mix} test set
- **ASR model settings:**
  16-layer Conformer encoder and 8-layer Transformer decoder (for CSE-SOT-Net)

| Detailed settings | Baseline | CSE-Net |
|---|---|---|
| Mixture (Mix.) encoder | 4 blocks | 4 blocks |
| Speaker differentiator (SpkrDiff.) encoder | 4 blocks | 4 blocks |
| Cross encoder | - | 2 blocks |
| Recognition (Rec.) encoder | 8 blocks | 6 blocks |

## Experimental results

- **Main results**

| ID | System | Librispeech | | LibrispeechMix-2mix | | | | |
|---|---|---|---|---|---|---|---|---|
| | | dev-clean | test-clean | Dev | Test (Overall) | Test (Conditional) (0, 0.2] | (0.2, 0.5] | (0.5, 1.0] |
| A1 | SIMO w/ PIT | **6.9** | **6.8** | 12.3 | 11.6 | 8.9 | 12.3 | 17.4 |
| A2 | SIMO w/ HEAT | 9.7 | 9.9 | **12.0** | **11.1** | **8.3** | **11.7** | **16.5** |
| B1 | SIMO w/ Joint-HEAT | 7.7 | 7.1 | 12.0 | 11.2 | 8.8 | 11.7 | 16.8 |
| B2 | CSE | **6.9** | **6.8** | **11.8** | **10.7** | **8.2** | **11.3** | **16.1** |
| C1 | - PPE | **6.9** | 6.8 | 11.9 | 11.0 | 8.4 | 11.7 | 16.8 |
| C2 | - mix. encoding | **6.9** | **6.7** | 11.9 | 10.9 | 8.3 | 11.7 | 16.4 |
| D1 | SOT | **4.2** | **5.4** | 9.5 | 9.4 | 7.3 | 9.9 | 13.9 |
| D2 | CSE-SOT | 4.5 | **5.5** | **8.1** | **8.4** | **7.2** | **8.3** | **12.0** |

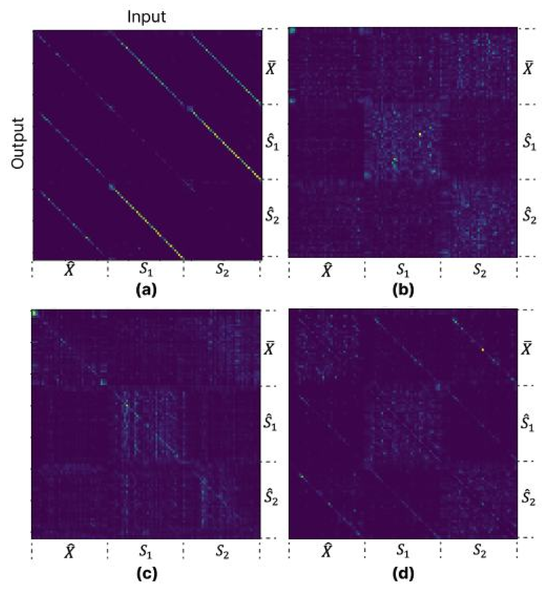
Loss function experiments, CSE experiments, and CSE-SOT experiments

- **Generalization results**

| System | Dev | Test (Overall) | Test (Conditional) (0,20] | (20,50] | (50,100] | OA-WER |
|---|---|---|---|---|---|---|
| SOT | **24.2** | **24.3** | **18.1** | **24.0** | **31.0** | **24.3** |
| CSE-SOT | **24.2** | 24.5 | **18.1** | **24.1** | 31.8 | 24.7 |

CSE-SOT network can generalize to 3-speaker case although there are only two branches in the CSE module.

- **Visualizations**




- Figures show attention matrices for 4 self-attention heads in cross-encoder
- Different attention heads attend distinct info:
  - Head **a**: frame-level info (lines)
  - Head **b**: partition-level info (blocks)
  - Head **c** and **d**: both frame/partition-level info
- Head (a) shows mutual attending between branches, implying branches swap
- Swapping of branches may serve a PIT-like function without the extra complexity

## Conclusions

- Discussed limitations of SIMO and SISO models for multi-talker speech recognition
- Proposed a cross-speaker encoding network (CSE-Net) consisting of a cross-encoder and joint-HEAT module
- CSE-Net can integrate with SOT, resulting in an integrated SIMO and SISO model
- Experiments validated the effectiveness of proposed models

Arxiv Link