The CW-XMU Systems For CHiME-6 Challenge

Member: Xuerui Yang, Yongyu Gao, Shi Qiu, Song Li, Qingyang Hong, Xuesong Liu, Lin Li, Dexin Liao, Hao Lu, Feng Tong, Qiuhan Guo, Huixiang Huang, Jiwei Li

Presented by: Song Li

CloudWalk Technology Co., Ltd., Shanghai, China

{yangxuerui, gaoyongyu, qiushi}@cloudwalk.cn

Xiamen University, Xiamen, China

{lilin, qyhong}@xmu.edu.cn

Outline

- Systems Overview
- Systems for Track1(RankA)
- Systems for Track1(RankB)
- Systems for Track2
- Results
- Conclusions

CHiME WORKSHOP

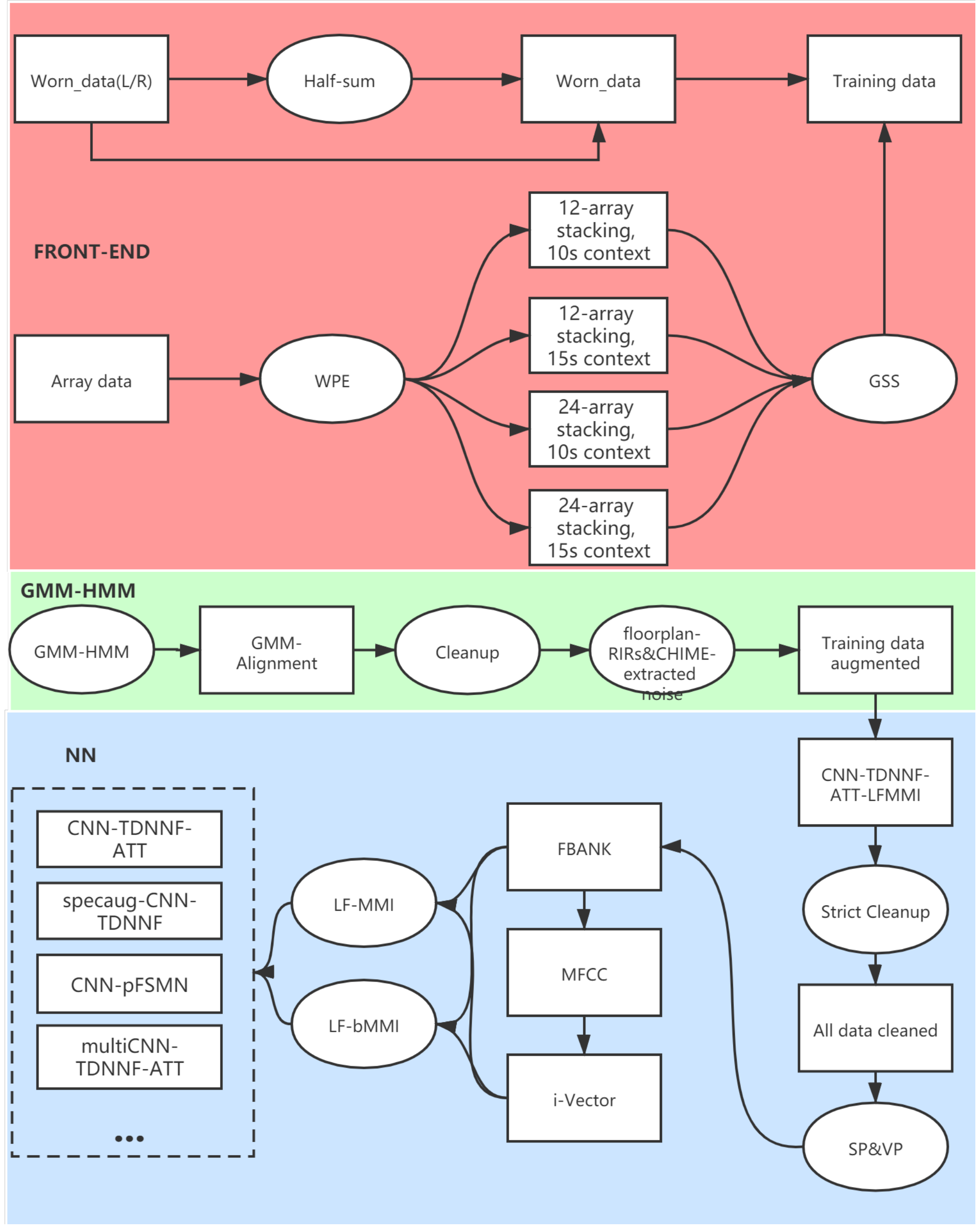
云从科技 CLOUDWALK

1.Systems Overview

- Track1(RankA)
  1. Front-end speech enhancement
  2. Acoustic model
  3. Alignment and Strict cleanup
  4. Chain-model tree leaves
  5. Model fusion
- Track1(RankB)
  1. Language model rescore
  2. End-to-End Model
- Track2
  1. Acoustic features
  2. Acoustic model
  3. Clustering algorithms
  4. Variational Bayesian refinement

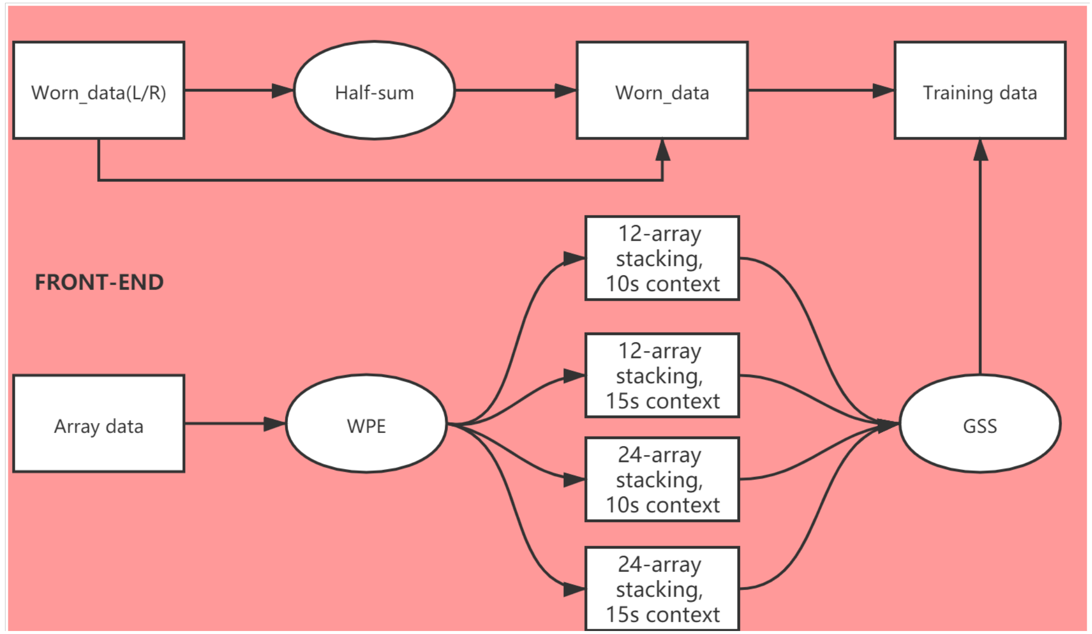
2. Systems for Track1(RankA)

- System is consist of Front-end, Gmm-Hmm and NN components



2. Systems for Track1(RankA)

2.1 Speech enhancement



- Weighted Predcition Error(WPE): Using nara_wpe tool and baseline Wpe configuration for derverberation
- Guided Source Separation (GSS): Baseline Gss with well-trained ASR model alignment
- Beamforming: Beamformit, Cgmm-MVDR
- Half-sum: Average the left and right channels of worn data

2. Systems for Track1(RankA)

2.2 Data Augmentation

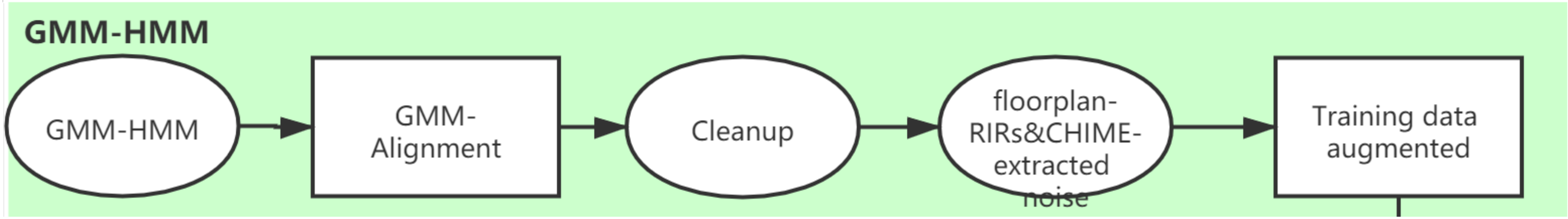
- Worn Data
  1. speed and volumn perturbation
  2. estimated noise injection
  3. simulated RIR: The RIR was generated according to the training data floorplan's configurations.

- Array Data: 24 and 12 micorphones GSS with various context length

- During training: SpecAugment

2. Systems for Track1(RankA)

2.3 GMM-HMM Training



- ➢ Strict cleanup : remove parts of speech, which has high WER
- ➢ Floorplan RIRs augmentation;
- ➢ Extract noise from CHiME-6 training data

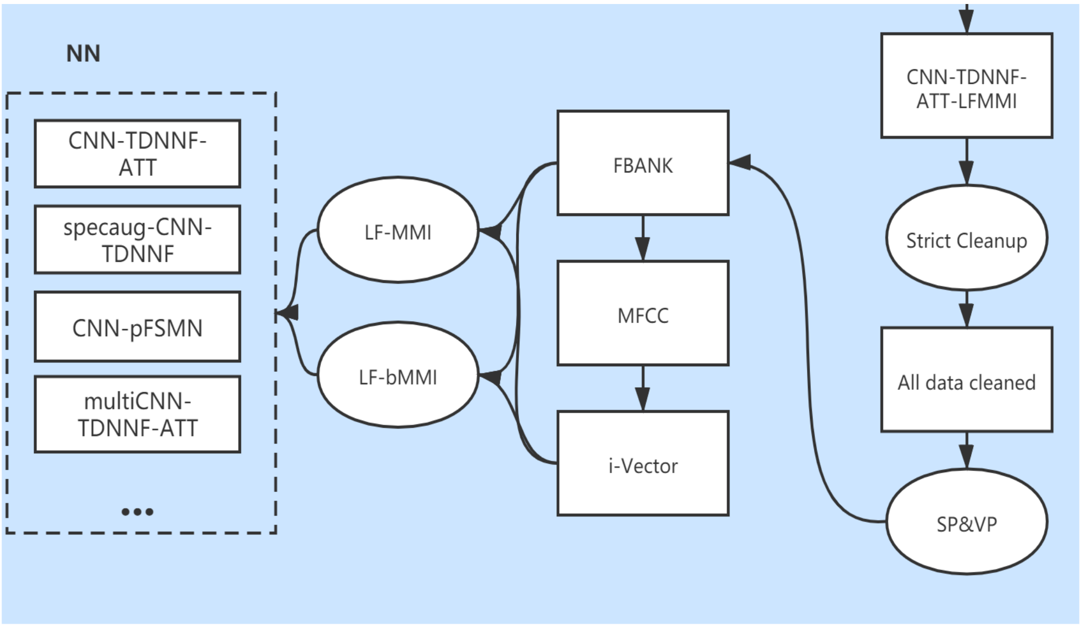
CHiME WORKSHOP

2. Systems for Track1(RankA)

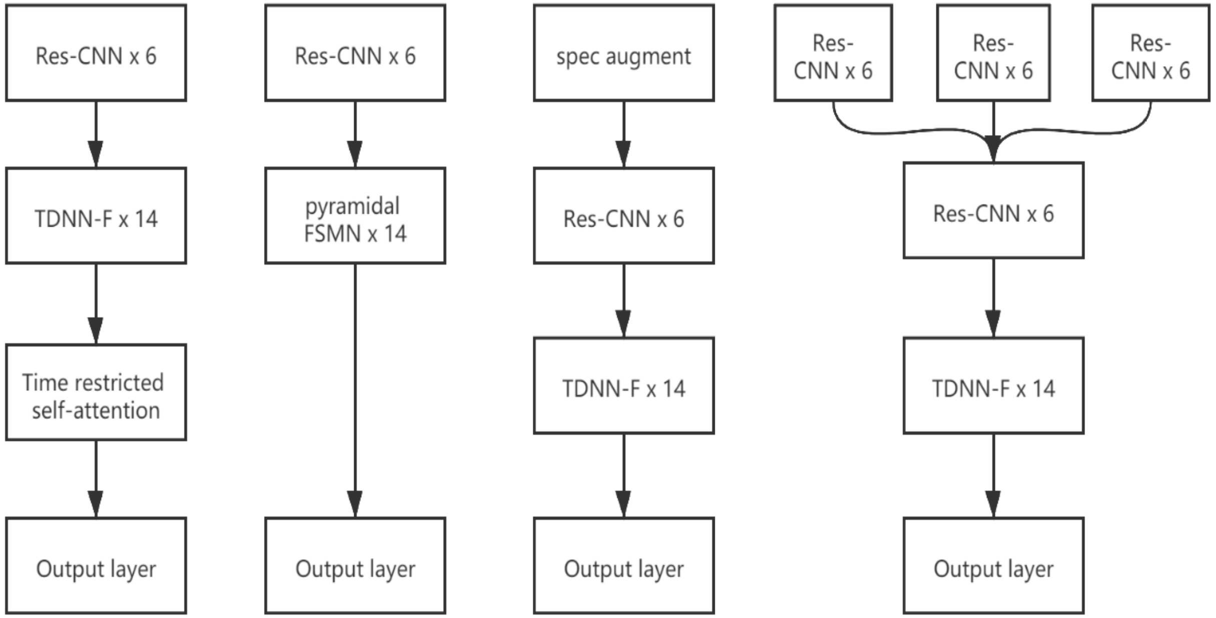
- 2.3 Chain Acoustic Model

- Acoustic Model

1.Res-cnn-tdnnf-self-attention
2.Res-cnn-fsmn
3.Spec-aug-cnn-tdnnf
4.Multi-cnn-tdnnf



Nnet training process



2. Systems for Track1(RankA)

2.3 Chain Model Training

- Neural-Network Alignment

  GMM-HMM

  ↓ Alignment

  Chain without subsampling

  ↓ Alignment

  Chain

- Chain-model tree leaves

  5000

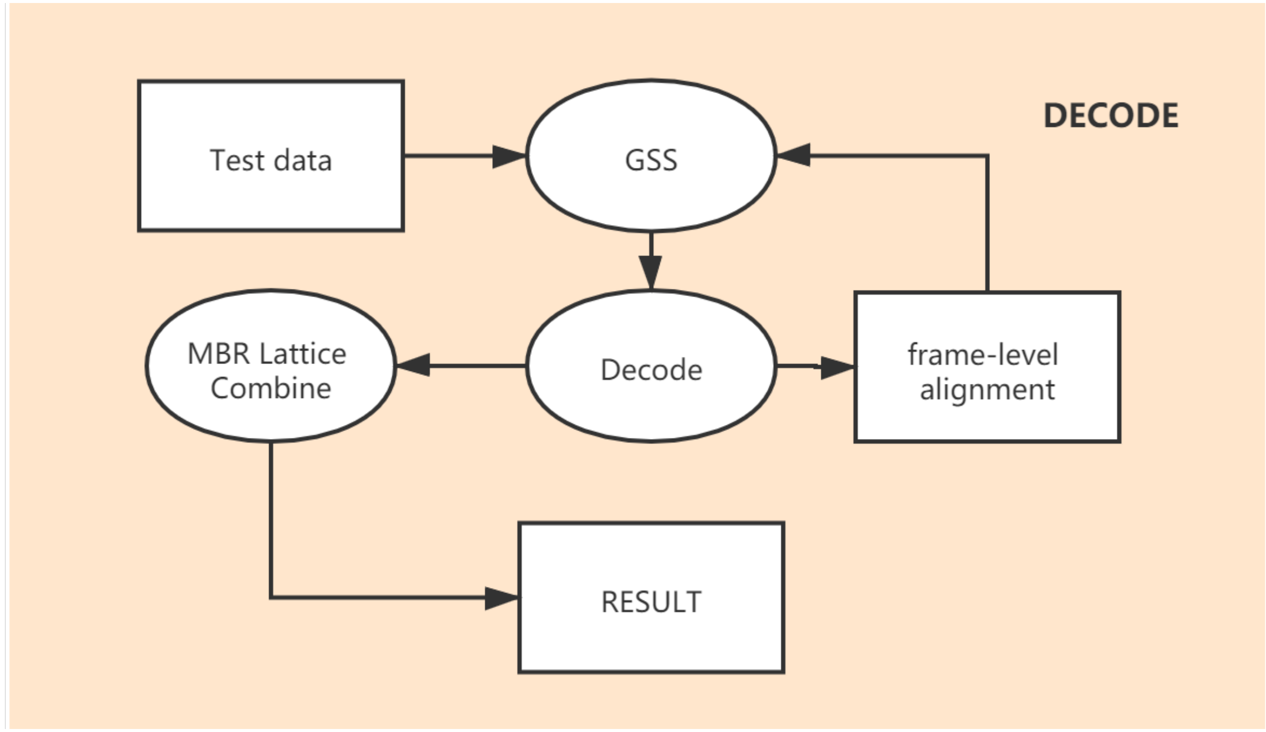
  ↓

  tree leaves

- Training criterion
  - LF-MMI;
  - LF-bMMI;

2. Systems for Track1(RankA)

2.4 Decode



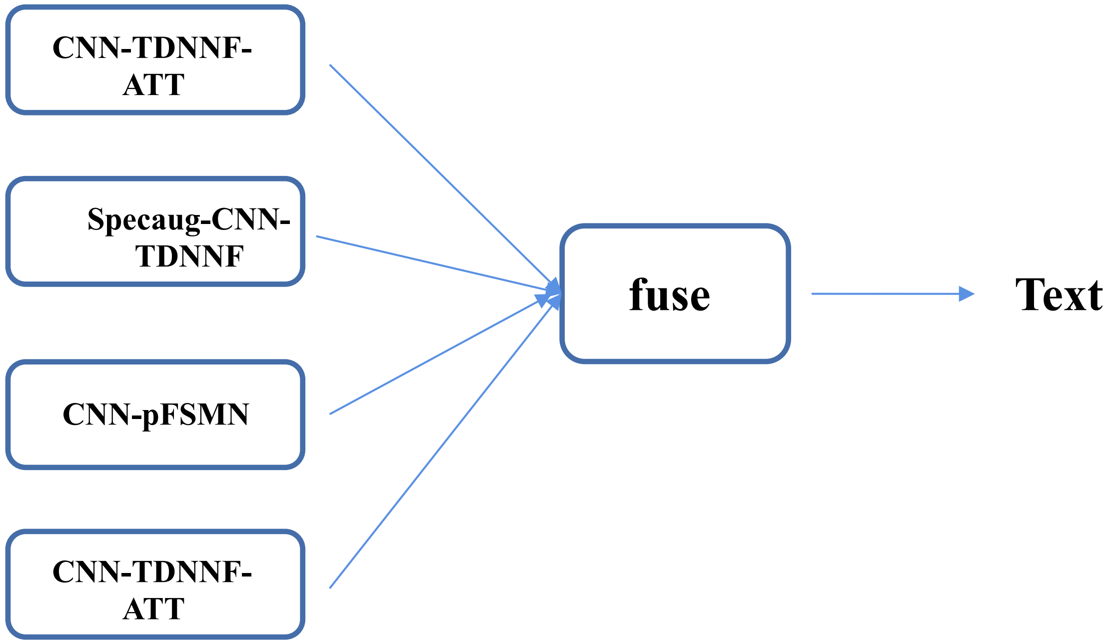
- Guided source separation
  - Alignment according to well-trained ASR model
  - with 10s context-length
  - Baseline WPE was used
  - Other dereverberation and beamforming was experimented

Model: Res-cnn-tdnnf-self-attention	
Baseline GSS, wpe and bf	**49.67**
15s context length GSS	49.64
CDR + baseline Gss	51.19
Baseline Gss + alignment	**48.46**
15s context length Gss + alignment	48.86

2. Systems for Track1(RankA)

2.4 Decode

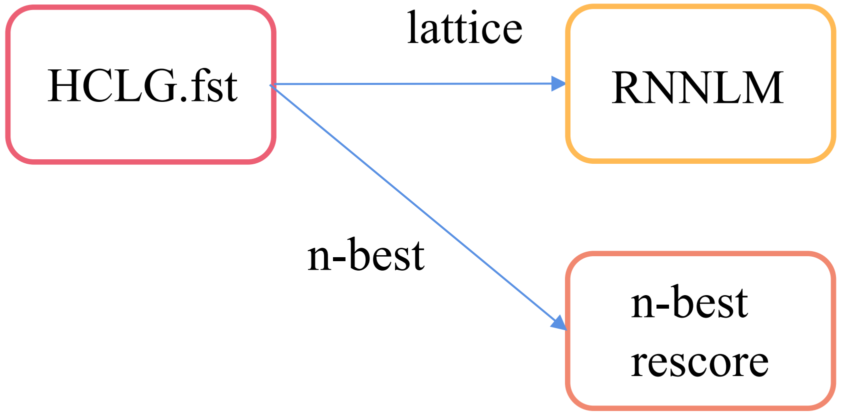
➢ Model fusion



- Minimum Bayesian Risk(MBR) Lattice Combine

3. Systems for Track1(RankB)

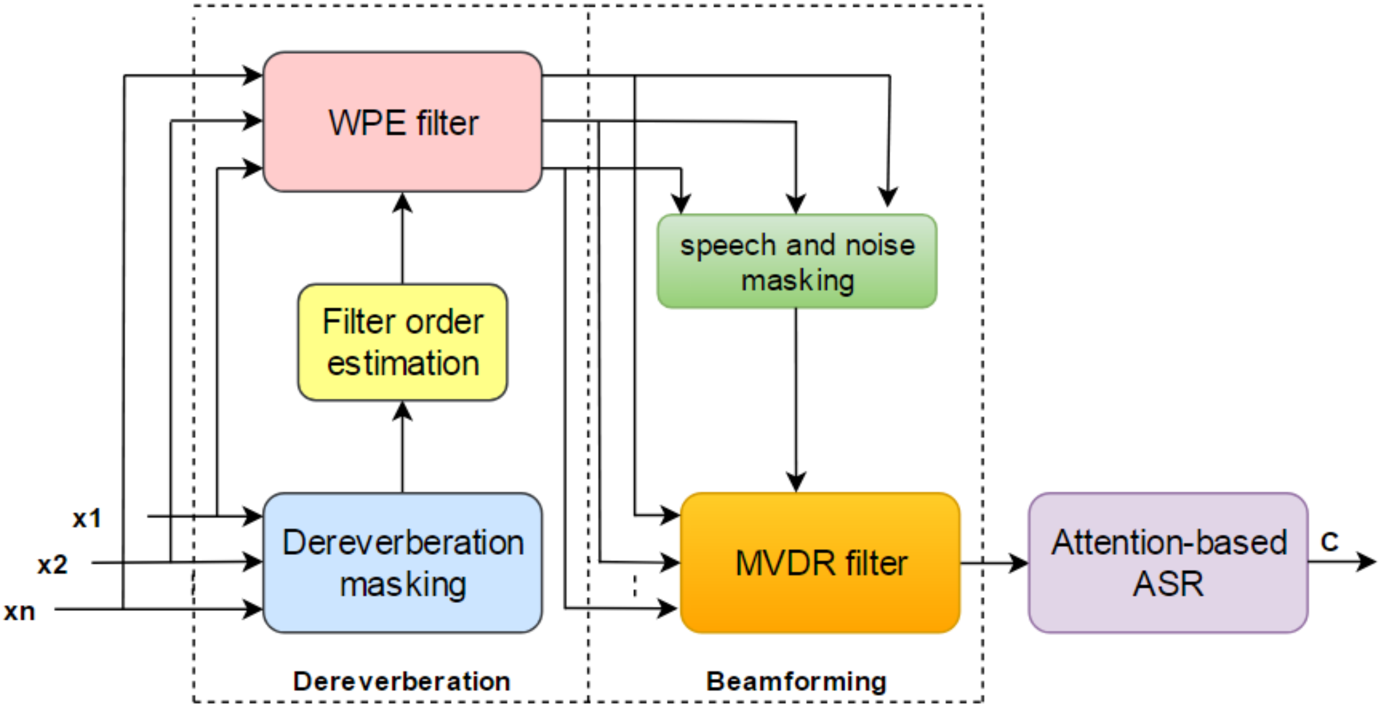
3.1 Language model rescore



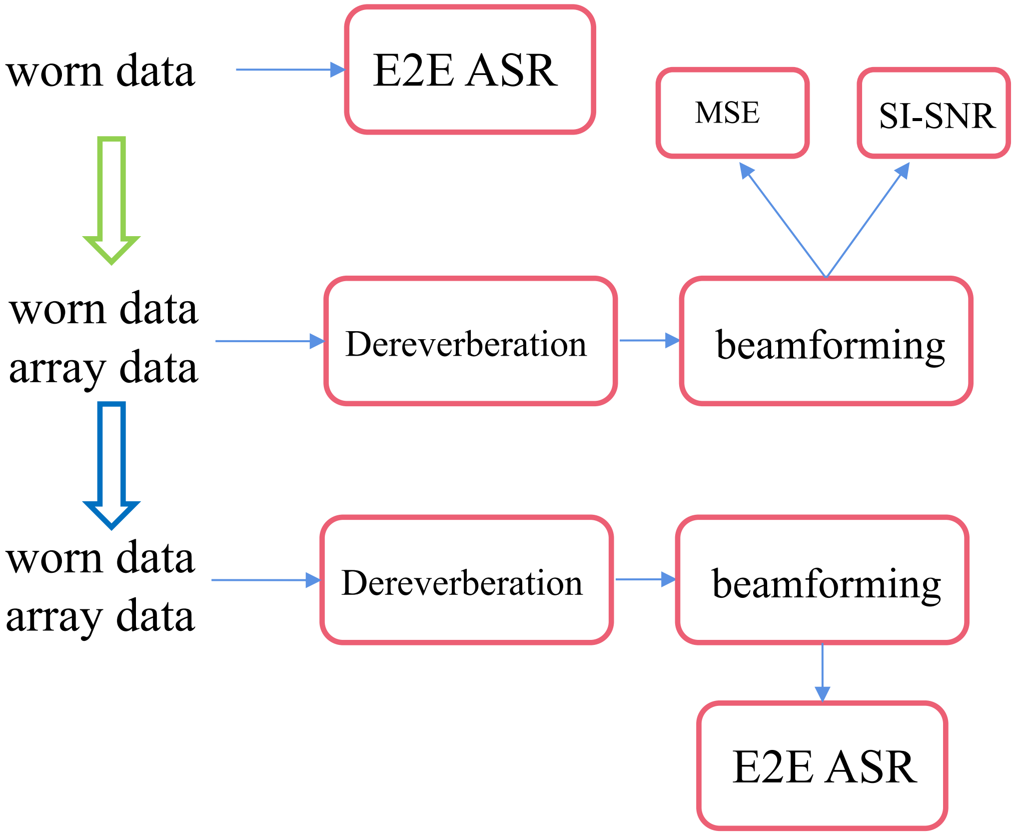
3. Systems for Track1(RankB)

3.1 End-to-End multi-channel ASR

- model structure

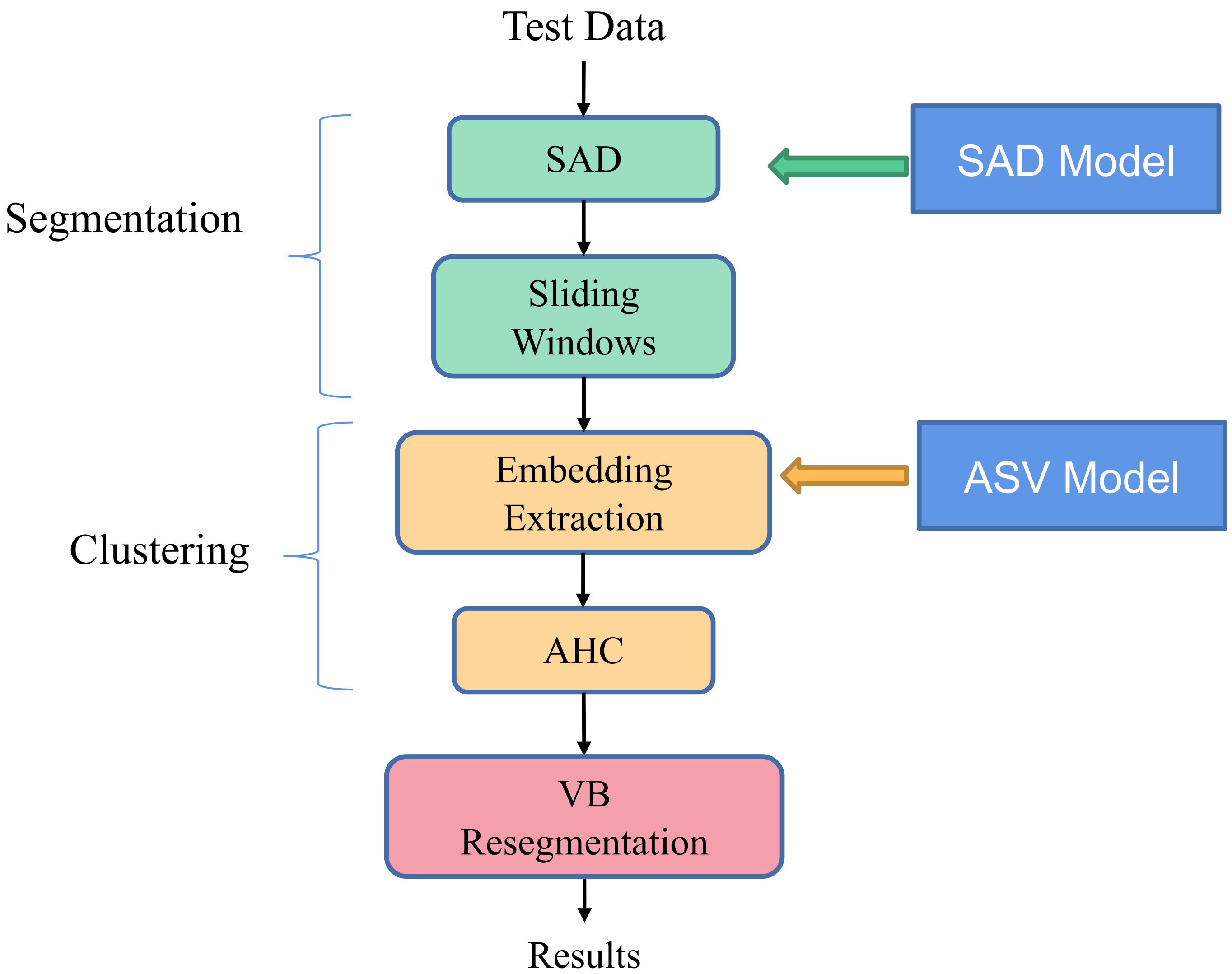


- Training process



4. Systems for Track2

- system structure



- **ASV Model**: F-TDNN
- **Acoustic features**:40-dimensional Fbank > 40-dimensional MFCC> 23-dimensional MFCC
- **Clustering Algorithms:** AHC >spectral clustering

5. Results

Track	Rank	Dev (WER %)	Eval (WER %)	
1	A	41.65	40.24	Hybrid systems
1	B	40.25	39.62	Hybrid systems
1	B	56.9	50.6	End-to-End

Table 1: Track 1 results

Baseline	Development Set			Evaluation Set		
	DER%	JER%	WER%	DER%	JER%	WER%
Category A	57.72	61.85	77.52	65.36	67.32	72.52

Table B: Track 2 category A results

3. Conclusions

- Using the frame-level alignment provided by ASR as the label of GSS can improve the performance of GSS.
- The chain model without subsampling provides better alignments.
- Removal of high WER speech during training can improve model performance.
- Different decision tree leaves can bring different performance to the model.
- SpecAugment is a kind of effective data augment method.
- Model fusion can improve recognition performance.
- Language model rescore is an effective post-processing method.
- End-to-end speech recognition in more difficult settings like reverberant, noisy, and far-field conditions, still lags behind.

Thank You!

Any Questions?