Technical Report

Zehua Liu
刘泽华

CNVSRC2024 Technical Report

Zehua Liu, Lantian Li, Dong Wang

Tsinghua University & Beijing University of Posts and Telecommunications

2024.08.16

NCMMSC-CNVSRC2024 Workshop, Urumqi, China

Outline

- Data and Tasks
- Baseline
- Technical Summary

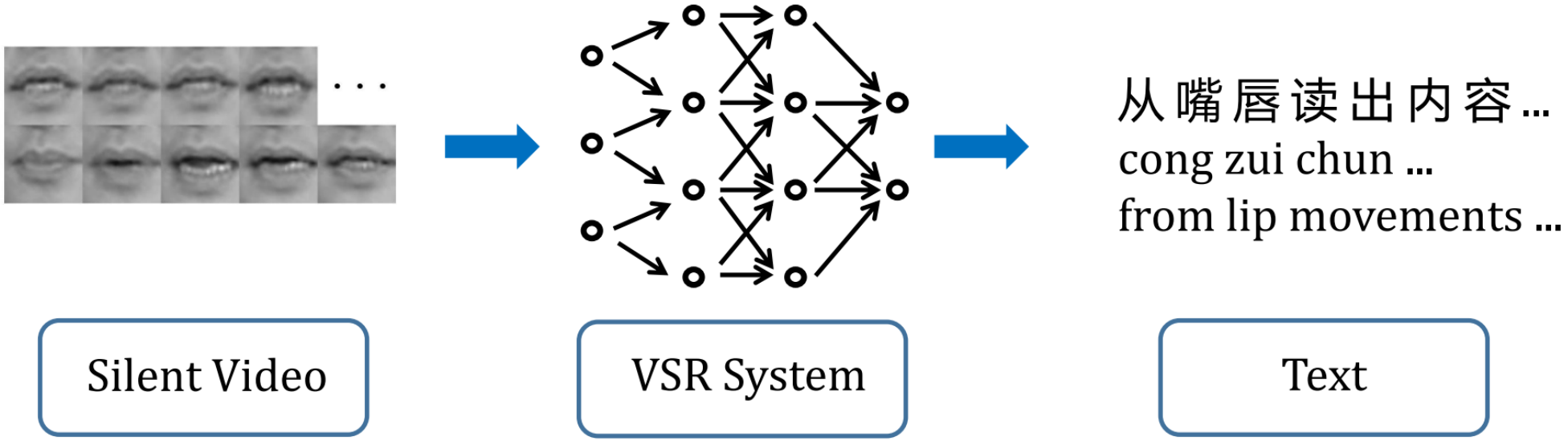
Outline

- **Data and Tasks**
- Baseline
- Technical Summary

Task Description

VSR(Visual Speech Recognition)

Definition



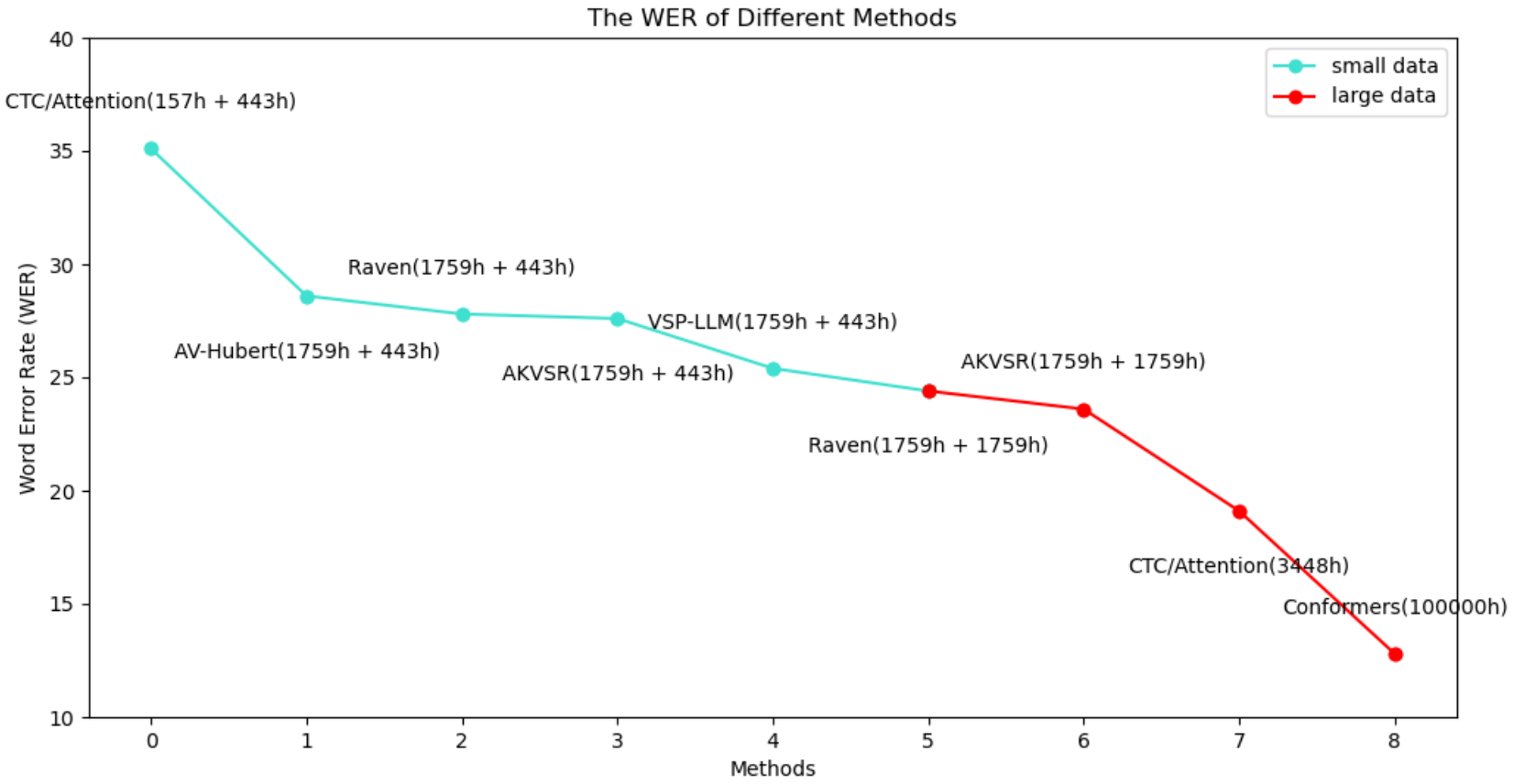
Performance measurement

$$\mathrm{CER} = \frac{\mathcal{N}_{\mathrm{Ins}} + \mathcal{N}_{\mathrm{Subs}} + \mathcal{N}_{\mathrm{Del}}}{\mathcal{N}_{\mathrm{Total}}} \times 100\%$$

Task Description

VSR(Visual Speech Recognition) in English

LRS3[1]

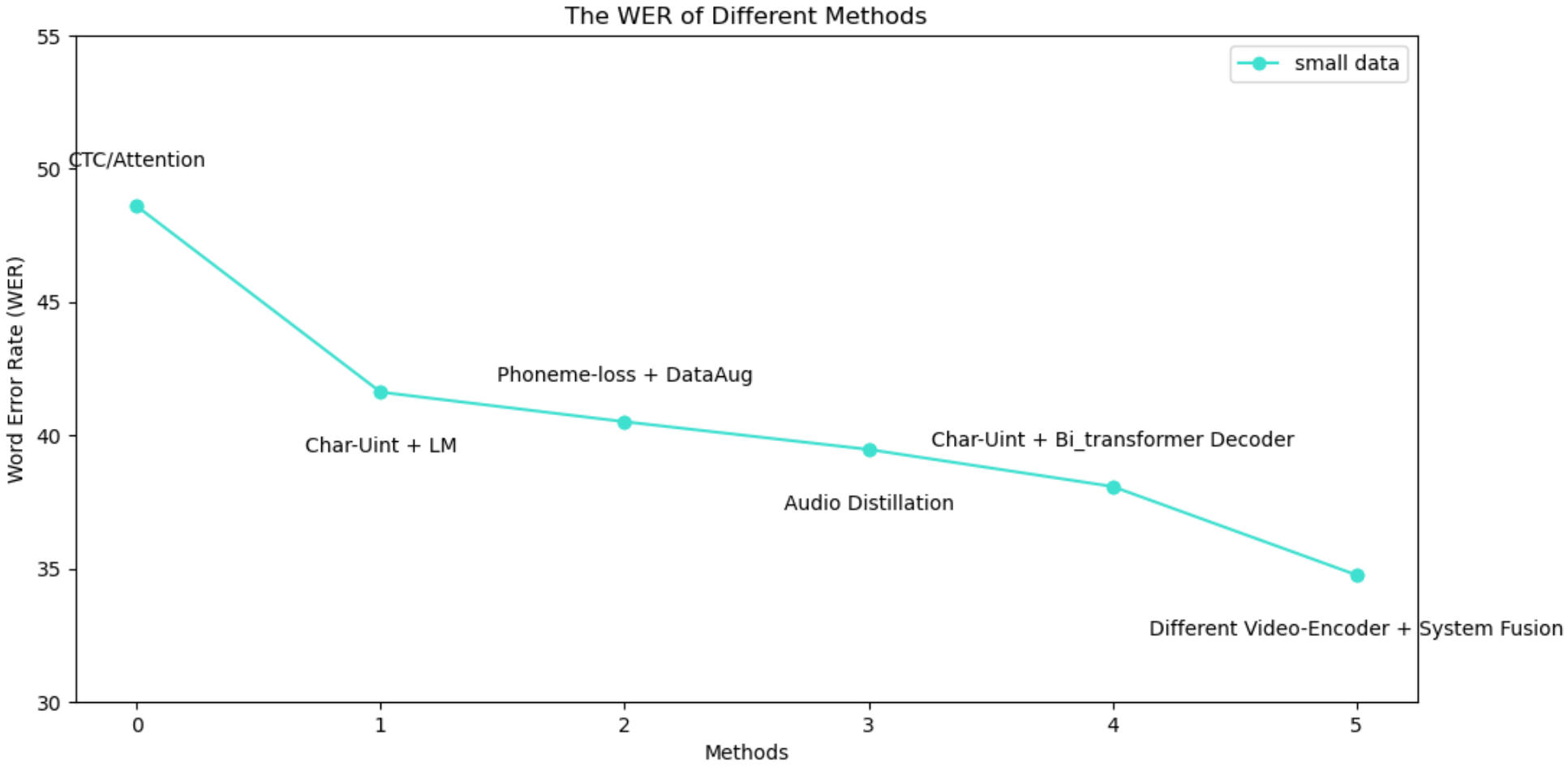


[1]T. Afouras, J. S. Chung, and A. Zisserman, “LRS3-TED: a large-scale dataset for visual speech recognition.”

Task Description

VSR(Visual Speech Recognition) in Chinese

CNVSRC 2023 Single-Speaker Fixed Track



Data

Dataset for CNVSRC 2024

- **CNCVS**
  - 2557 Speakers
  - 300 Hours Audio-video paired data
- **CN-CVS2-P1**
  - Additional Dataset for open track
  - 200 Hours Audio-video paired data
- **CNVSRC – Single**
  - 1 Speaker
  - 100 Hours Audio-video paired data ,collect from Internet
- **CNVSRC – Multi**
  - 43 Speakers
  - 1 Hour per Speaker Audio-video paired Data

Task Description

T1 Single-Speaker VSR

- Fixed Track
  - ONLY CN-CVS and CNVSRC-Single.Dev is allowed for training/tuning ALL the components of the system.
  - This track is designed to compare different techniques under the SAME data resource.
- Open Track
  - ANY data sources can be used for developing ALL the components of the system.
  - This track is designed to examine the performance Frontier of the present technologies.

	Fixed Track	Open Track
T1: Single-speaker VSR	CN-CVS, CNVSRC-Single.Dev	No constraint (e.g. CN-CVS2-P1)
T2: Multi-speaker VSR	CN-CVS, CNVSRC-Multi.Dev	No constraint (e.g. CN-CVS2-P1)

Task Description

T2 Multi-Speaker VSR

- Fixed Track
  - ONLY CN-CVS and CNVSRC-Multi.Dev is allowed for training/tuning ALL the components of the system.
  - This track is designed to compare different techniques under the SAME data resource.
- Open Track
  - ANY data sources can be used for developing ALL the components of the system.
  - This track is designed to examine the performance Frontier of the present technologies.

	Fixed Track	Open Track
T1: Single-speaker VSR	CN-CVS, CNVSRC-Single.Dev	No constraint (e.g. CN-CVS2-P1)
T2: Multi-speaker VSR	CN-CVS, CNVSRC-Multi.Dev	No constraint (e.g. CN-CVS2-P1)

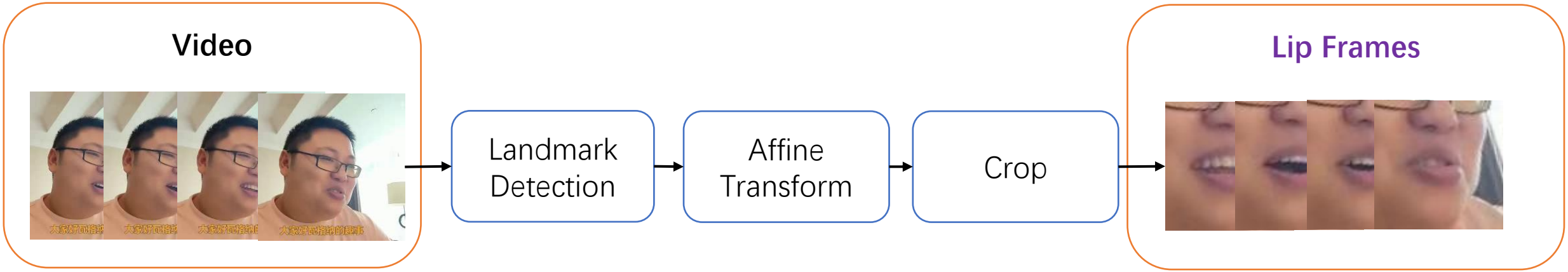
Outline

- Data and Tasks
- **Baseline**
- Technical Summary

Baseline

Data processing

Video Preprocess

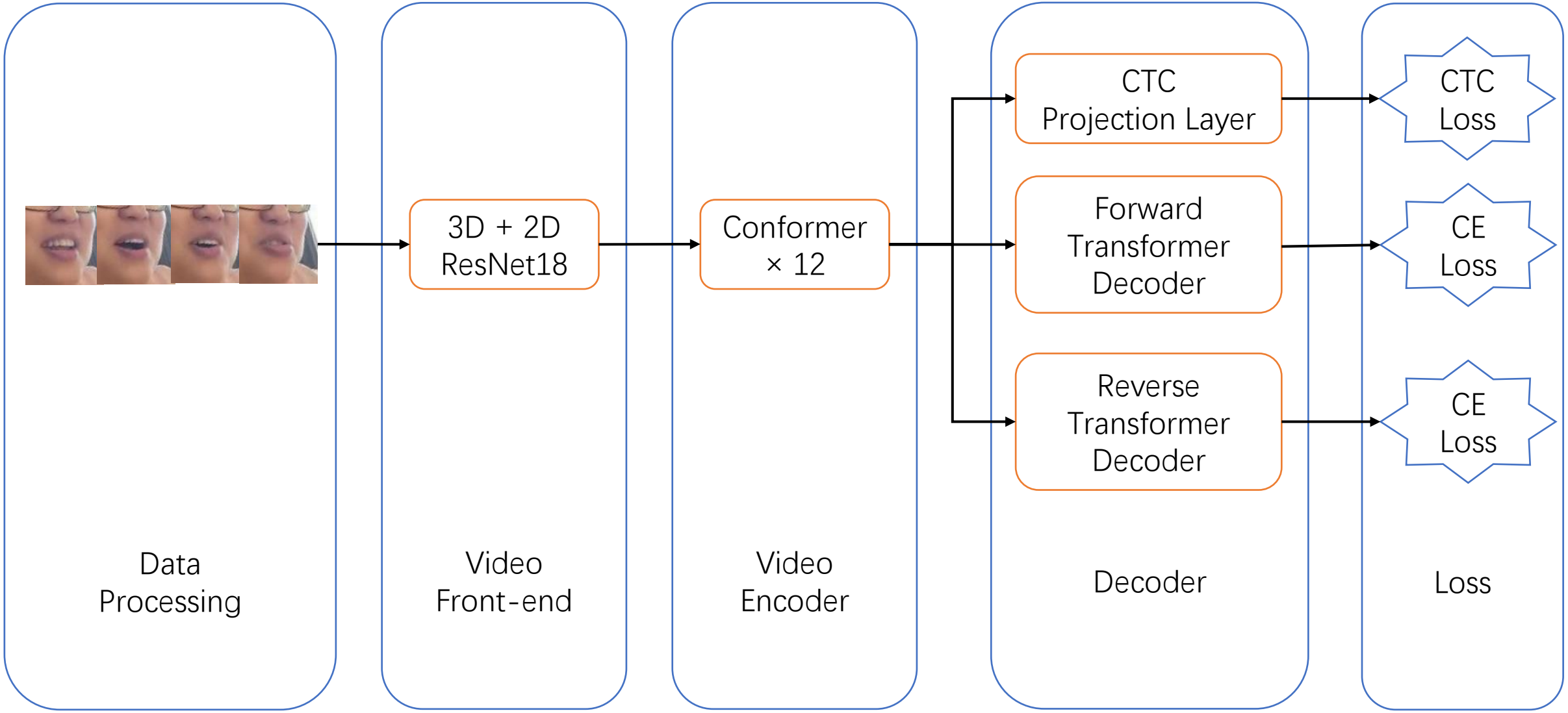


Model Input

The input of the VSR model is **Lip Frames**

Baseline

Model Structure



Baseline

CNVSRC2024 vs CNVSRC 2023

CNVSRC 2024 Improvement

- Reverse Transformer Decoder
- Model units from subword units (5,904 units) into char units (4,468 units)

Result

	T1: Single-speaker VSR		T2: Multi-speaker VSR	
	CNVSRC 2024	CNVSRC 2023	CNVSRC 2024	CNVSRC 2023
Valid	41.22%	48.57%	52.42%	58.77%
Eval	39.66%	48.60%	52.20%	58.37%

Outline

- Data and Tasks
- Baseline
- **Technical Summary**

Technical Summary

Components	Methods
Data processing	Face Detection, Face Alignment, Multi-scale Lip Region Extraction
Data Augmentation	Speed Perturbation, Adaptive Time Masking, Random Crop, Flip, Color Transformation, Image Size Crop
Visual Front-end	Resnet3D, Enhanced Resnet3D
Encoder	Conformer, Branchformer, E-Branchformer
Decoder	Transformer Decoder, Reverse Transformer Decoder, S4D Decoder, Reverse S4D Decoder
Loss function	CTC/Attention Loss, KLDivLoss
Training strategy	Pretrain, Fine-tune
System fusion	Score-level Average

Technical Highlights

Components	Methods
Data processing	Face Detection, Face Alignment, Multi-scale Lip Region Extraction
Data Augmentation	Speed Perturbation, Adaptive Time Masking, Random Crop, Flip, Color Transformation, **Image Size Crop**
Visual Front-end	Resnet3D, **Enhanced Resnet3D**
Encoder	Conformer, Branchformer, E-Branchformer
Decoder	Transformer Decoder, Reverse Transformer Decoder, **S4D Decoder, Reverse S4D Decoder**
Loss function	CTC/Attention Loss, KLDivLoss
Training strategy	**Pretrain**, Fine-tune
System fusion	**Score-level Average**

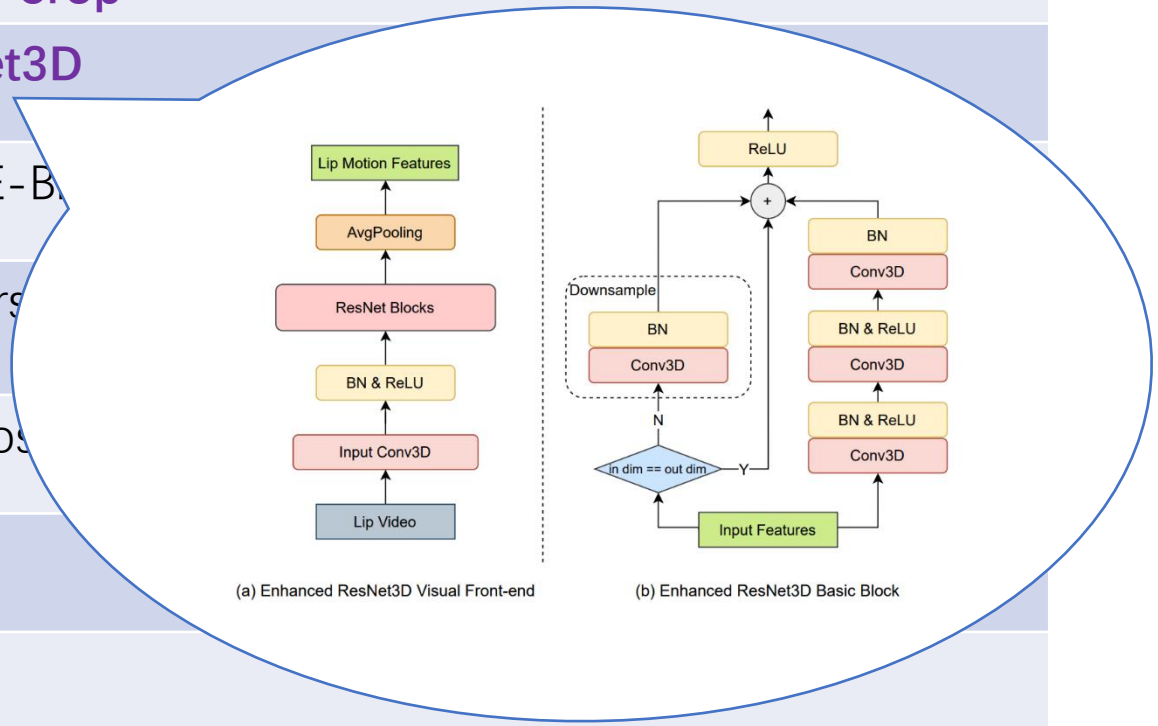
Data Augmentation

Components	Methods
Data processing	Face Detection, Face Alignment, Multi-scale Lip Region Extraction
Data Augmentation	Speed Perturbation, Adaptive Time Masking, Random Crop, Flip, Color Transformation, **Image Size Crop**
Visual Front-end	Resnet3D, **Enhanced Resnet3D**
Encoder	Conformer, Branchformer, E-Branchfo
Decoder	Transformer Decoder, Reverse Tran **S4D Decoder**
Loss function	CTC/Attention Loss, KLDivLoss
Training strategy	**Pretrain**, Fine-tune
System fusion	**Score-level Average**

System	Visual Frontend	Encoder	Decoder	Crop	T1.Dev
M1	ResNet3D	E-Branchformer	Transformer	80	41.76
M2	ResNet3D	Conformer	Transformer	96	39.27
M3	ResNet3D	Branchformer	Transformer	96	38.86
M4	ResNet3D	E-Branchformer	Transformer	96	38.39
M5	Enhanced ResNet3D	E-Branchformer	Transformer	96	38.30
M6	ResNet3D	E-Branchformer	Transformer	112	38.13
M7	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	112	36.60
M8	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	128	36.48
R1	ROVER **Eval** of M1∼M8				
R2	ROVER **R1** and **Eval.FT** of M1∼M8				

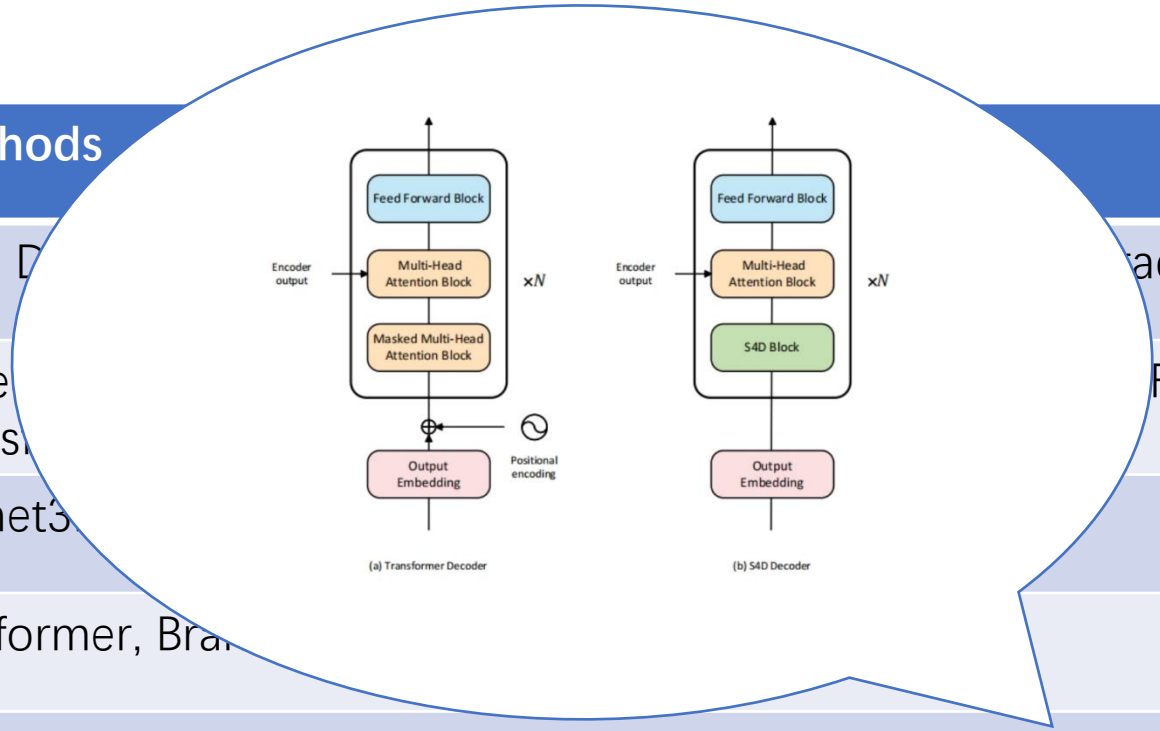
Visual Front-end

Components	Methods
Data processing	Face Detection, Face Alignment, Multi-scale Lip Region Extraction
Data Augmentation	Speed Perturbation, Adaptive Time Masking, Random Crop, Flip, Color Transformation, **Image Size Crop**
Visual Front-end	Resnet3D, **Enhanced Resnet3D**
Encoder	Conformer, Branchformer, E-B
Decoder	Transformer Decoder, Revers **S4D Decoder**
Loss function	CTC/Attention Loss, KLDivLos
Training strategy	**Pretrain**, Fine-tune
System fusion	**Score-level Average**



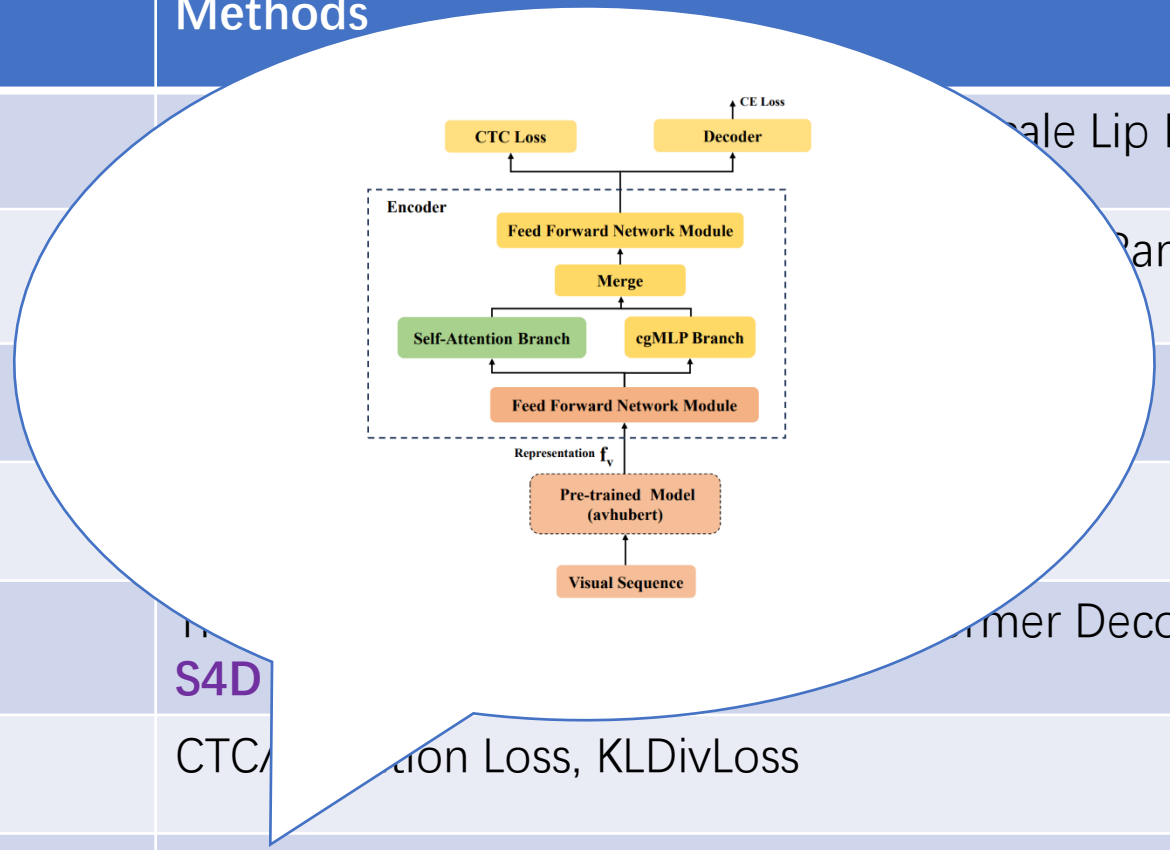
Decoder

Components	Methods
Data processing	Face D… …action
Data Augmentation	Spee… Trans… …Flip, Color
Visual Front-end	Resnet3…
Encoder	Conformer, Bra…
Decoder	Transformer Decoder, Reverse Transformer Decoder, **S4D Decoder, Reverse S4D Decoder**
Loss function	CTC/Attention Loss, KLDivLoss
Training strategy	**Pretrain**, Fine-tune
System fusion	**Score-level Average**



Training Strategy

Components	Methods
Data processing	ale Lip Region Extraction
Data Augmentation	Random Crop, Flip, Color
Visual Front-end	
Encoder	
Decoder	mer Decoder, **S4D Decoder, Reverse S4D**
Loss function	CTC/ tion Loss, KLDivLoss
Training strategy	**Pretrain**, Fine-tune
System fusion	**Score-level Average**



System fusion

Components	Methods
Data processing	Face Detection, Face Alignment, Multi-scale Lip Region Extraction
Data Augmentation	Speed P… …lip, Color Tra…
Visual Front-end	
Encoder	
Decoder	…erse
Loss function	CTC/…
Training strategy	**Pretrain**, Fin…
System fusion	**Score-level Average**

System	Visual Frontend	Encoder	Decoder	Crop	T1.Dev	T1.Eval	T1.Eval.FT	T2.Dev	T2.Eval	T2.Eval.FT
M1	ResNet3D	E-Branchformer	Transformer	80	41.76	-	-	48.50	-	-
M2	ResNet3D	Conformer	Transformer	96	39.27	-	-	46.21	-	-
M3	ResNet3D	Branchformer	Transformer	96	38.86	-	-	46.49	-	-
M4	ResNet3D	E-Branchformer	Transformer	96	38.39	-	-	45.97	-	-
M5	Enhanced ResNet3D	E-Branchformer	Transformer	96	38.30	-	-	45.79	45.43	41.75
M6	ResNet3D	E-Branchformer	Transformer	112	38.13	-	35.91	44.88	44.56	40.37
M7	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	112	36.60	36.79	34.56	43.29	43.13	38.73
M8	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	128	36.48	37.04	**34.27**	43.23	42.99	**37.93**
R1	ROVER **Eval** of M1∼M8					32.7199			38.3602	
R2	ROVER **R1** and **Eval.FT** of M1∼M8					**30.4679**			**34.2955**	


CNVSRC 2024
Chinese Continuous Visual Speech Recognition Challenge

Many Thanks!