CNVSRC 2023 Technical Report

Chen Chen, Lantian Li, Dong Wang

Tsinghua University & Beijing University of Posts and Telecommunications

2023.12.09

NCMMSC-CNVSRC 2023 Workshop, Suzhou, China

OUTLINE

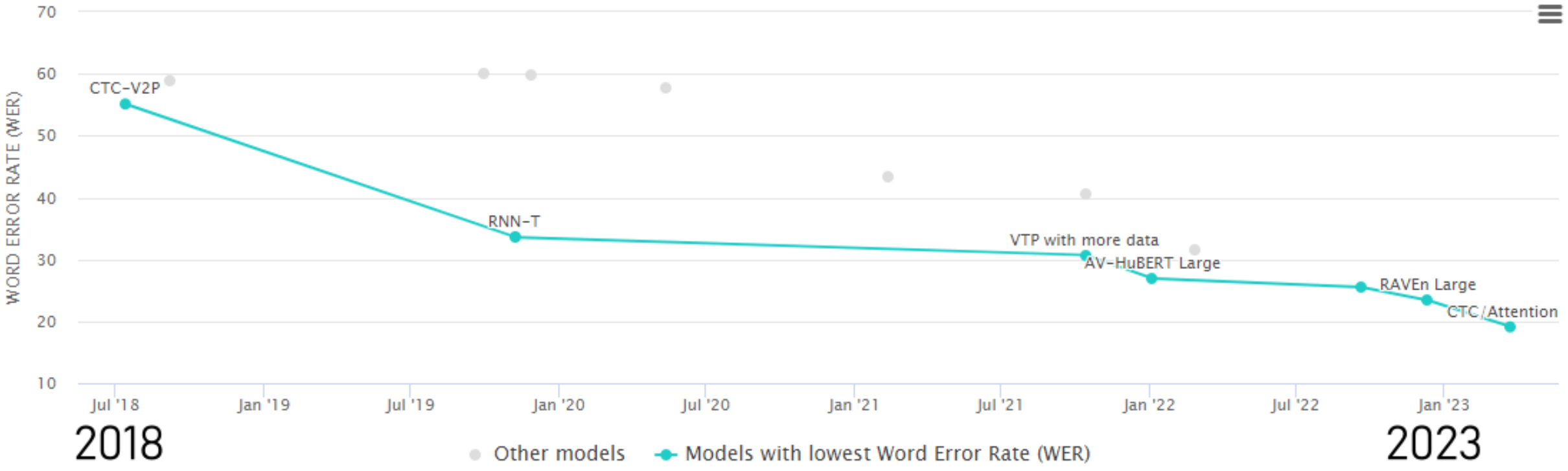
- Data and Tasks
- Baseline
- Technical Summary

OUTLINE

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

The Origin of CN-CVS Dataset

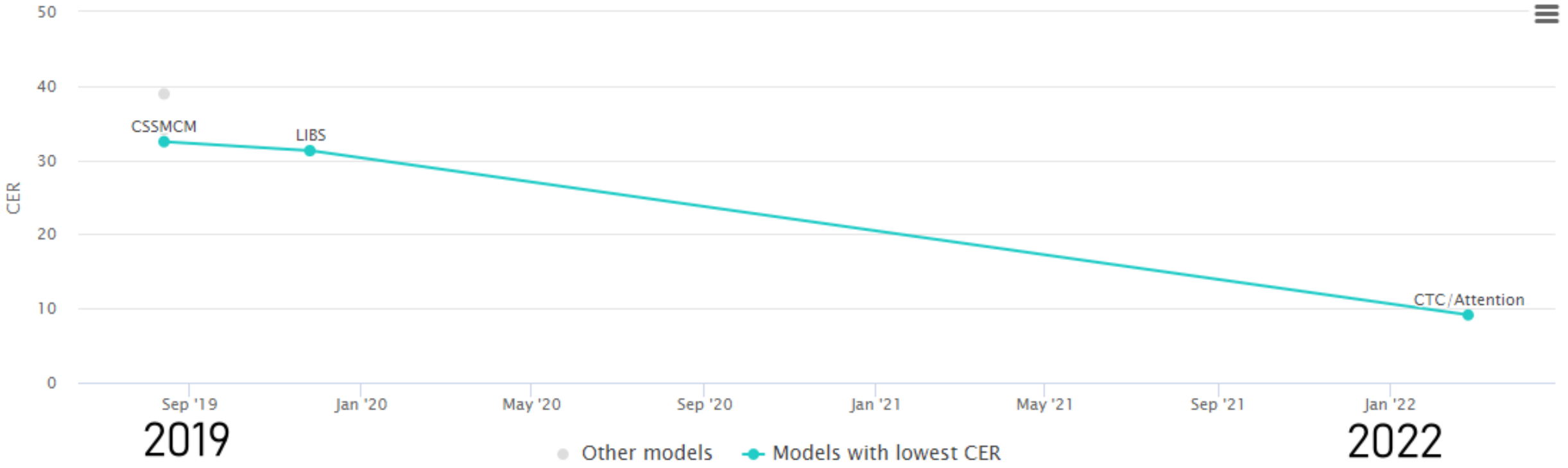
- The lack of Chinese Audio-video data constrain the development of Chinese VSR
  - English VSR: Auto-AVSR[1] reach 19.1% WER on LRS3[2] **in the wild**.



[1] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels," ICASSP 2023
[2] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition."
https://paperswithcode.com/sota/lipreading-on-lrs3-ted

The Origin of CN-CVS Dataset

- The lack of Chinese Audio-video data constrain the development of Chinese VSR
  - Chinese VSR: [3] reach 9.1% WER on CMLR[4] **in the lab**.



[3] P. Ma, S. Petridis, and M. Pantic, “Visual Speech Recognition for Multiple Languages in the Wild.”
[4] Y. Zhao, R. Xu, and M. Song, “A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading.”
https://paperswithcode.com/sota/lipreading-on-cmlr

The Origin of CN-CVS Dataset

- The lack of Chinese Audio-video data constrain the development of Chinese VSR

数据集名称	文献	语言	内容类型	数据来源	词汇量	说话人数量	句子数量	时长	摄像机角度
GRID	Cooke et al. (2006)	英语	语法模式	录制朗读	51	33	33000	27h	0°
TCD-TIMIT	Harte et al. (2015)	英语	句	录制朗读	5954	62	6913	≈20h	0°, 30°
Lip2Wav	Prajwal et al. (2020)	英语	句	讲课节目	≈5k/spk	5	-	≈120h	自然角度
LRW	Chung et al. (2017a)	英语	词	电视新闻	500	-	≈539000	173h	自然角度
LRS	Chung et al. (2017b)	英语	句	电视新闻	≈17k	-	118116	75.5h	自然角度
LRS2	Afouras et al. (2019)	英语	句	电视新闻	≈60k	-	≈145000	224.5h	自然角度
LRS3	Afouras et al. (2019)	英语	句	演讲节目	≈70k	≈10k	≈165000	475h	自然角度
VoxCeleb1	Nagrani et al. (2017)	英语	句	网络视频	-	1251	153516	352h	自然角度
VoxCeleb2	Chung et al. (2018)	多语言	句	网络视频	-	6112	1128246	2442h	自然角度
AVSpeech	Ephrat et al. (2018)	多语言	句	演讲授课	-	≈150k	-	≈4700h	自然角度
CAS-VSR-W1k	Yang et al. (2019)	汉语	词	电视节目	1k	>2000	718018	≈140h	自然角度
CMLR	Zhao et al. (2019)	汉语	句	电视新闻	3517	11	102076	≈88h	0°
CN-CVS/News	Chen et al. (2023)	汉语	句	电视新闻	-	28	13016	34.6h	0°
CN-CVS/Speech	Chen et al. (2023)	汉语	句	演讲节目	-	2529	193245	273.4h	自然角度

Data Collection Pipeline

1 Programs selection

新闻类节目
单人公开演讲
科普演讲
...

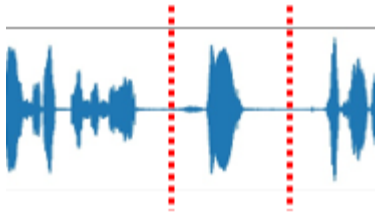
2 Videos download

bilibili SELF 格致论道 造就 一刻talks 席

3 Shot detection, VAD and Face Tracking

Shot detection
FFmpeg

VAD
pydub



Face tracking
dlib

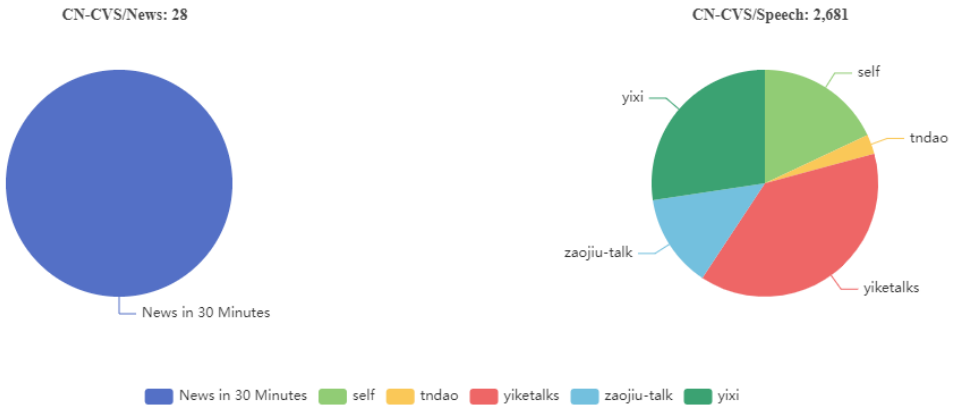
4 Mouth-Speech Sync

Synchronization
SyncNet

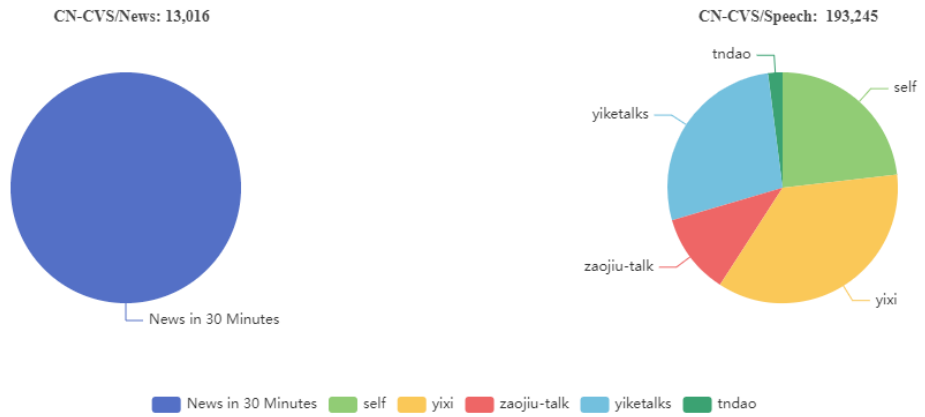
5 Human check

Data Profile

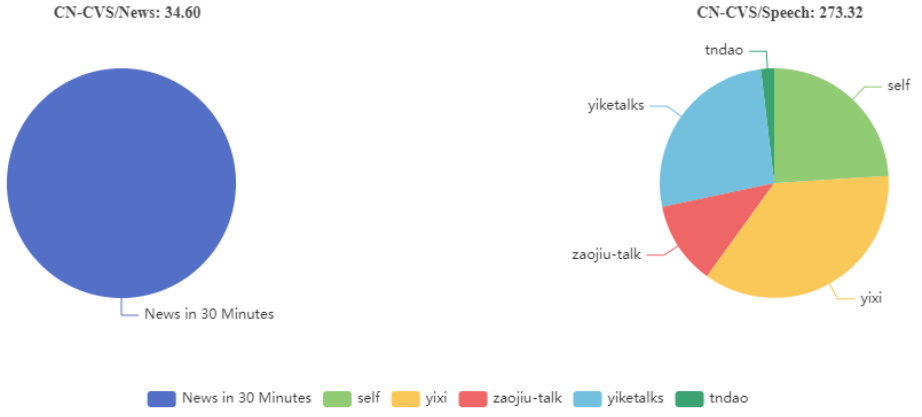
2,557 Speakers



206,261 Utters



~300 Hours



[5] C. Chen, D. Wang, and T. F. Zheng, “CN-CVS: A Mandarin Audio-Visual Dataset for Large Vocabulary Continuous Visual to Speech Synthesis,” ICASSP 2023

Data Profile

- Additional datasets for CNVSRC 2023
  - **CNVSRC-Single**
    - 1 Speaker
    - 100 Hours Audio-video paired data
  - **CNVSRC-Multi**
    - 43 Speakers
    - 1 Hour per Speaker Audio-video paired Data

	CNVSRC-Single		CNVSRC-Multi	
DataSet	Dev	Eval	Dev	Eval
# Videos	25,947	2,881	20,450	10,269
# Hours	94.00	8.41	29.24	14.49

Task Description – 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
T2: Multi-speaker VSR	CN-CVS, CNVSRC-Multi.Dev	No constraint

Task Description – 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

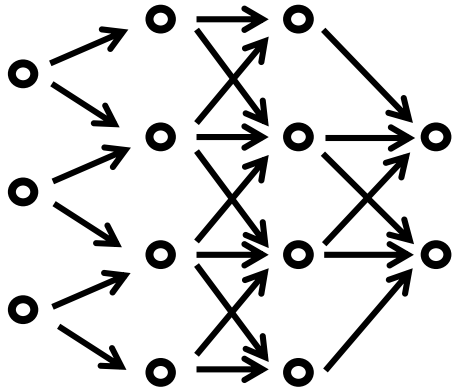
- **ANY** data sources can be used for developing **ALL** the components of the system.
- This track is designed to examine the performance **Frontier** of the present technologies.

	Fixed Track	Open Track
T1: Single-speaker VSR	CN-CVS, CNVSRC-Single.Dev	No constraint
T2: Multi-speaker VSR	CN-CVS, CNVSRC-Multi.Dev	No constraint

Task Description – VSR

- Definition

Silent Video



VSR System

从嘴唇读出内容...
cong zui chun ...
from lip movements ...

Text

- Performance measurement

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

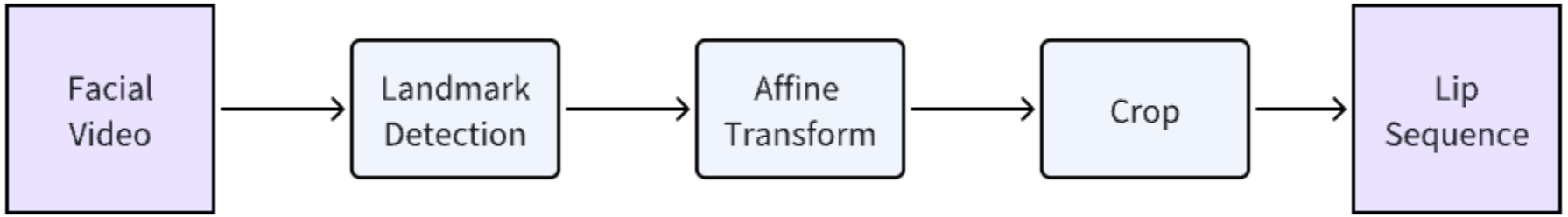
OUTLINE

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

Baseline

Data processing

- **Video Data**

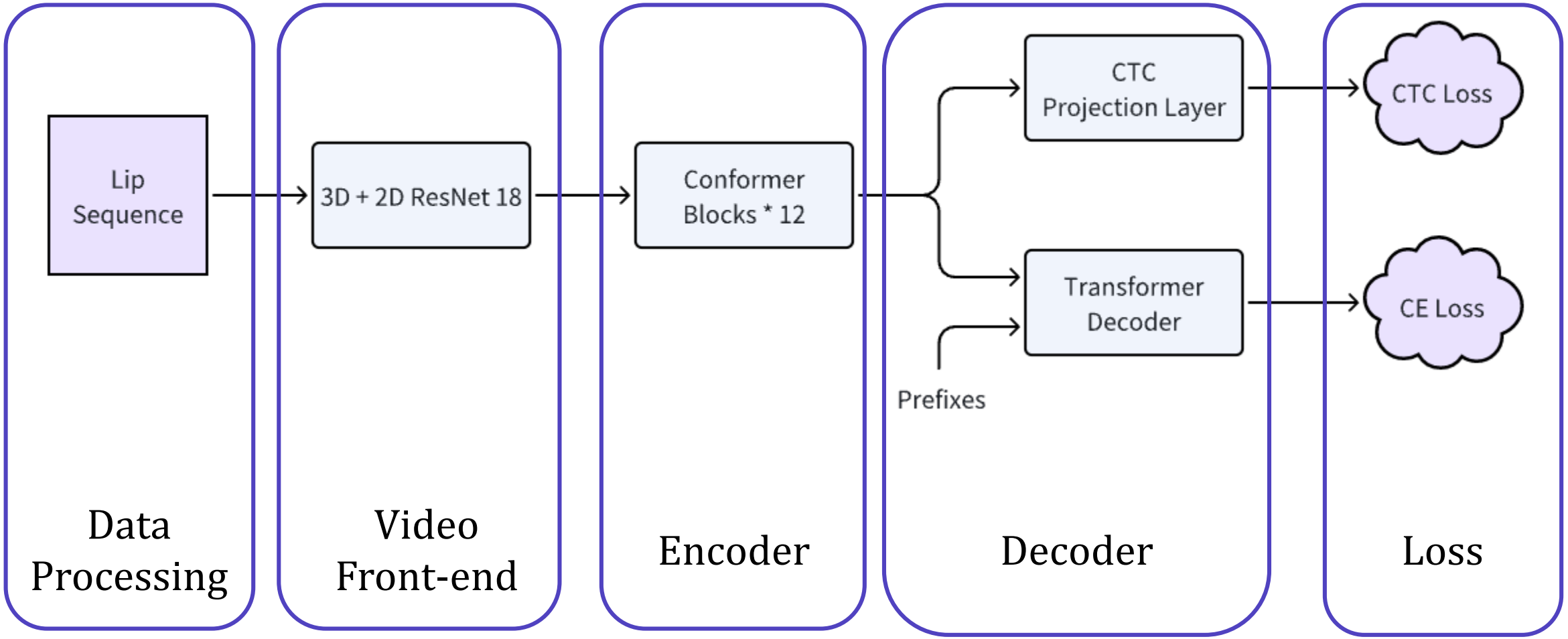


- **Text Data**
  - SentencePiece *

*https://pypi.org/project/sentencepiece/

Baseline

Model Structure



OUTLINE

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

Representative Techniques (6 Teams)

Components	Methods
Data processing	Face Detection, Face Alignment, Multi-scale Lip Region Extraction
Data Augmentation	Speed perturbation, Adaptive Time Masking, Random Erase, Flip, Generated Facial Video
Video Front-end	3D+2D ResNet18, ResNet3D
Encoder	Conformer, Branchformer, E-Branchformer
Decoder	Transformer Decoder, Bi-Transformer Decoder
Auxiliary design	BPE/Character as Modeling Unit, Phoneme-level prediction task
Loss function	CTC/Attention Loss, CTC Loss in shallow layers, Cross Modality Similarity Loss
Training strategy	Pretrain + Fine-tune, ASR Knowledge Distillation
Language Model	RNNLM, Transformer LM
System fusion	Score-level average

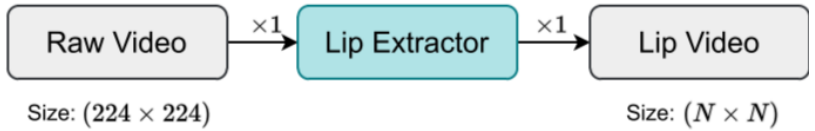
Technical Highlights

Components	Methods
Data processing	Face Detection, Face Alignment, **Multi-scale Lip Region Extraction**
Data Augmentation	**Speed perturbation**, Adaptive Time Masking, Random Erase, Flip, Generated Facial Video
Video Front-end	3D+2D ResNet18, ResNet3D
Encoder	Conformer, Branchformer, E-Branchformer
Decoder	Transformer Decoder, Bi-Transformer Decoder
Auxiliary design	**Character as Modeling Unit**, Phoneme-level prediction task
Loss function	CTC/Attention Loss, **CTC Loss in shallow layers**, Cross Modality Similarity Loss
Training strategy	Pretrain + Fine-tune, **ASR Knowledge Distillation**
Language Model	RNNLM, Transformer LM
System fusion	Score-level average

Data processing

Components	Methods
Data processing	Face Detection, Face Alignment, **Multi-scale Lip Region Extraction**
Data Augmentation	**Speed perturbation**, Adaptive Tim sking, Random Erase, Flip, Generated Facial Video
Video Front-end	3D+2D Res
Encoder	
Decoder	
Auxiliary	k
Loss func	ty Similarity Loss
Training s	
Language Mo	
System fusion	

T237



System	Encoder	Crop	SP	T1.Dev	T1.Eval	T2.Dev	T2.Eval
Baseline[6]	Conf	96	✗	48.57	48.60	58.77	58.37
M1	Conf	96	✓	39.43	39.99	46.08	45.73
M2	Branch	96	✓	39.00	39.36	46.63	46.37
M3	E-Branch	96	✓	38.59	38.61	46.26	45.80
M4	E-Branch	48	✗	46.88	45.81	55.58	55.51
M5	E-Branch	64	✗	44.40	43.59	53.64	52.98
M6	E-Branch	80	✗	42.95	42.26	50.77	50.38
M7	E-Branch	96	✗	40.56	40.42	47.16	46.53
M8	E-Branch	112	✗	38.46	38.95	45.17	44.87
ROVER	-	-	-	**34.47**	**34.76**	**41.39**	**41.06**

Data Augmentation

Components	Methods
Data processing	Face Detection, Face Align… Region Extraction
Data Augmentation	**Speed perturbation**, A… Random Erase, Flip, Generated Facial Video
Video Front-end	3D+2D ResNet18, ResNet3D
Encoder	Conformer, Branchformer, E-Branchformer
Decoder	Transformer Decoder, Bi-Transformer Decoder
Auxiliary design	Character as Modeling Unit, Phoneme-level prediction task
Loss function	CTC/Attention Loss, CTC Loss in shallow layers, Cross Modality Similarity Loss
Training strategy	Pretrain + Fine-tune, ASR Knowledge Distillation
Language Model	RNNLM, Transformer LM
System fusion	Score-level average

T237, T238

Auxiliary design

Components	Methods
Data processing	Face Detection, Face Alignment,
Data Augmentation	Speed perturbation, Adaptive Ti
Video Front-end	3D+2D ResNet18, ResNet3D
Encoder	Conforme -Bra
Decoder	Transformer D oder, Bi-Transformer Decoder
Auxiliary design	**Character as Modeling Unit**, Phoneme-level prediction task
Loss function	
Training strategy	
Language Model	
System fusion	

T266, T267

System	Model	CER (%)
M1	Proposed system	40.46
M2	M1 - RNNLM	40.62
M3	M2 - char unit	42.36
M4	M3 - Bi-transformer decoder	43.19
M5	M4 - Inter CTC residual module	48.57

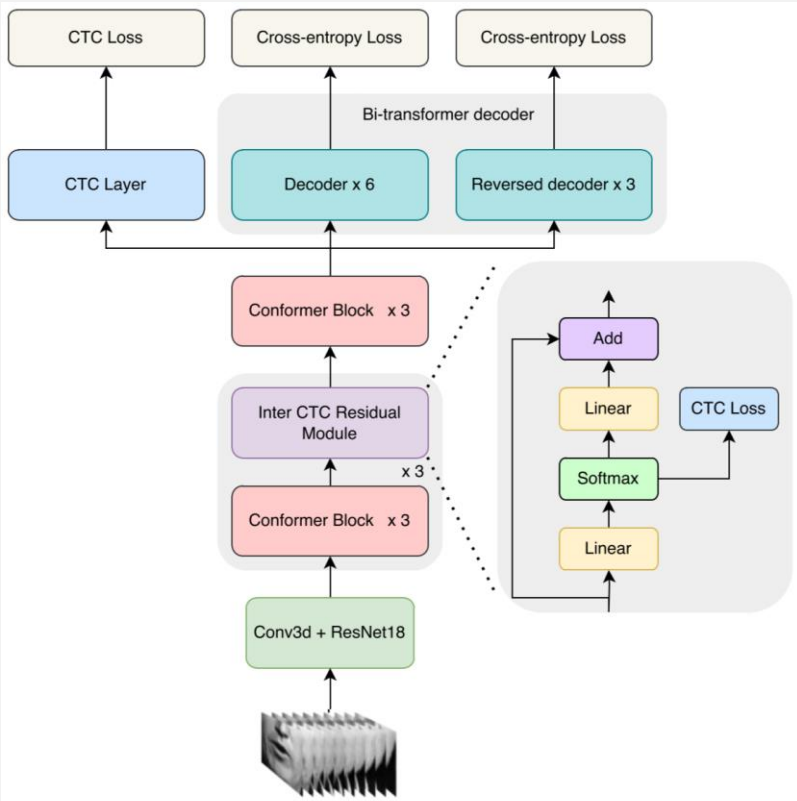
method	training set(hours)	finetuning set(hours)		single-speaker dev CER	multi-speakers dev CER
		T1	T2		
baseline	287	83.7	18	48.57%	58.77%
char model unit	287	83.7	18	43.59%	56.77%
+video aug	287	166	36	42.30%	54.74%
+RNN LM				42.18%	
+transformer LM				42.16%	
+model fusion				**41.50%**	

Loss function

Components	Methods
Data proc	ip Regio
Data Aug	and
Video Fr	
Encoder	
Decoder	
Auxiliary	edict
Loss function	CTC/Attention Loss, **CTC Loss in shallow layers**, C
Training strategy	Pretrain + Fine-tune, ASR Knowledge Distillation
Language Model	RNNLM, Transformer LM
System fusion	Score-level average

T266

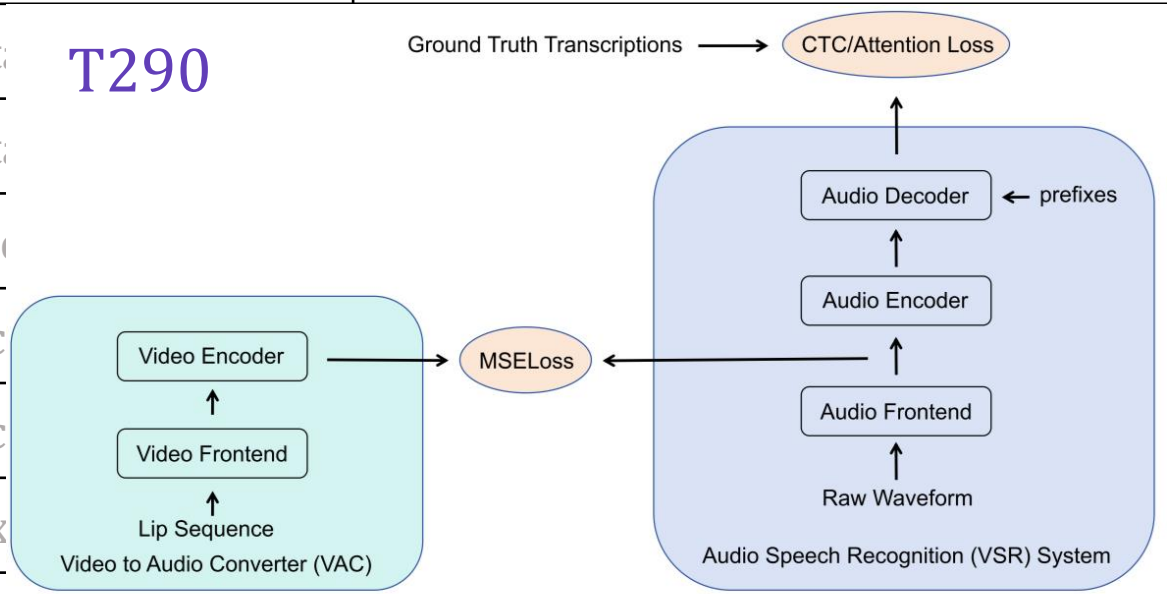
System	Model	CER (%)
M1	Proposed system	40.46
M2	M1 - RNNLM	40.62
M3	M2 - char unit	42.36
M4	M3 - Bi-transformer decoder	43.19
M5	M4 - Inter CTC residual module	48.57



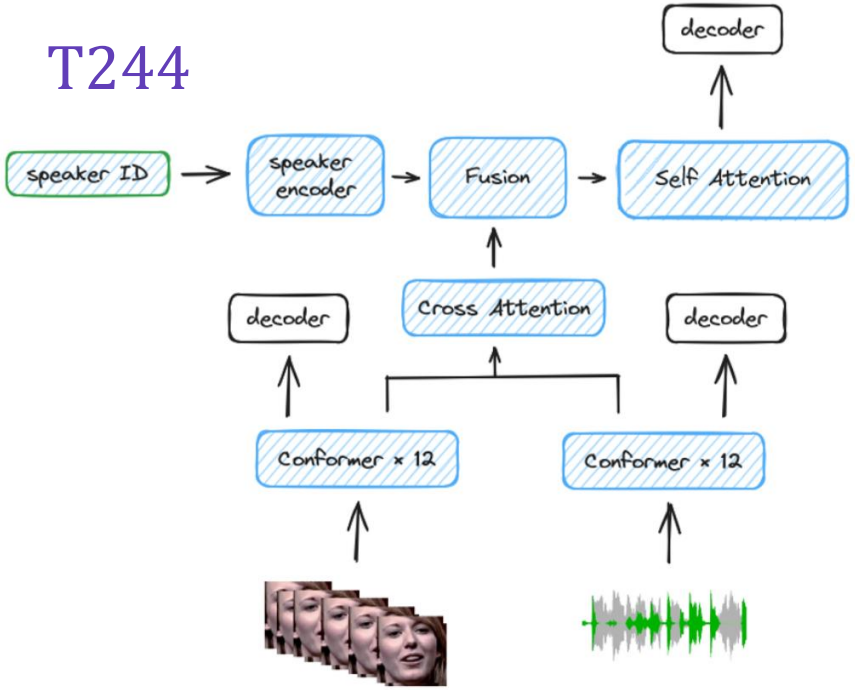
Training strategy

Components	Methods
Data	scale Lip Regi
Data	sking, Rando
Vide	
Enc	mer
Dec	Decoder
Aux	evel predicti
Loss function	w layers, Cro
Training strategy	Pretrain + Fine-tune, **ASR Knowledge Distillation**
Language Model	RNNLM, Transformer LM
System fusion	Score-level average

T290



T244



Many Thanks !

National Conference on Man-Machine Speech Communication 2023 NCMMSC

CIPS

CCF

TSINGHUA UNIVERSITY

BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

海天瑞声

Speechhome