The VII-BUPT (T238) System for T1. Fixed Track

Zehua Liu, Xiaolou Li, Yiwei Sun, Zhenzhen Pan, Li Guo

Beijing University of Posts and Telecommunications

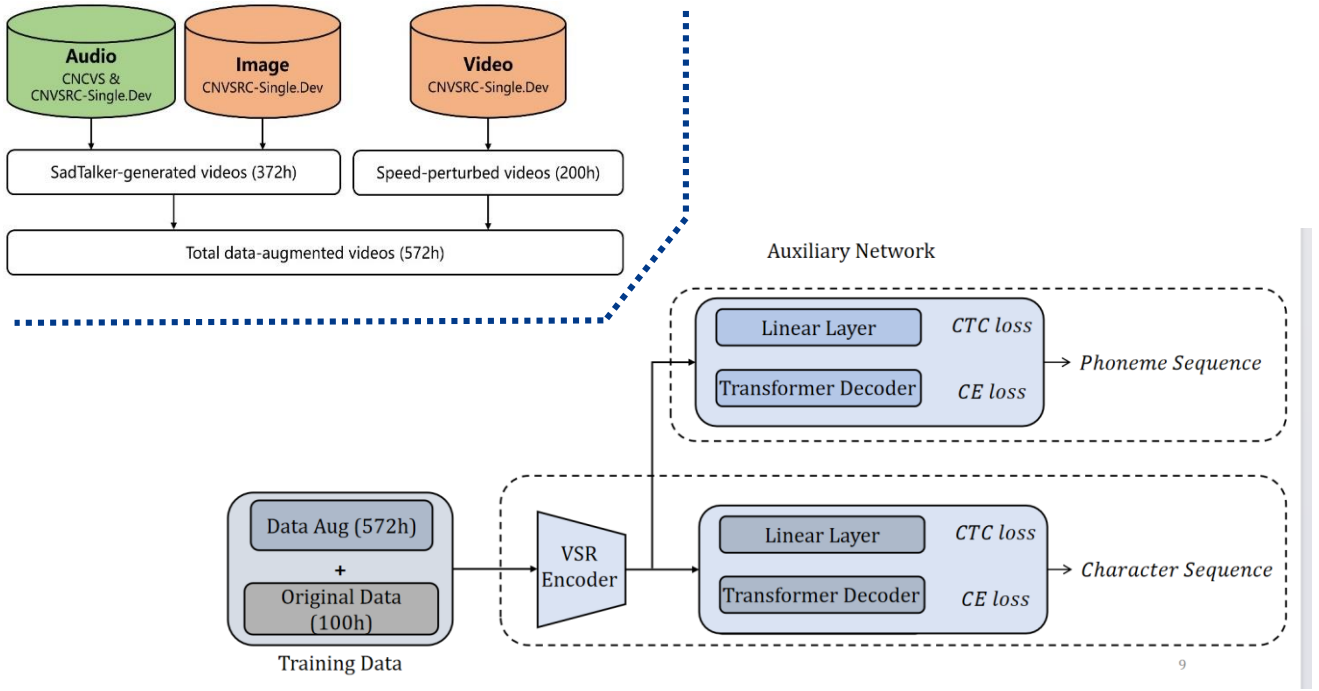
Overview T238 VII-BUPT System

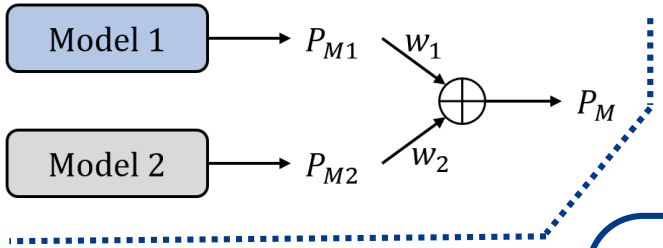
Data Augmentation

Character-Phoneme VSR Model

System Fusion

Language Model





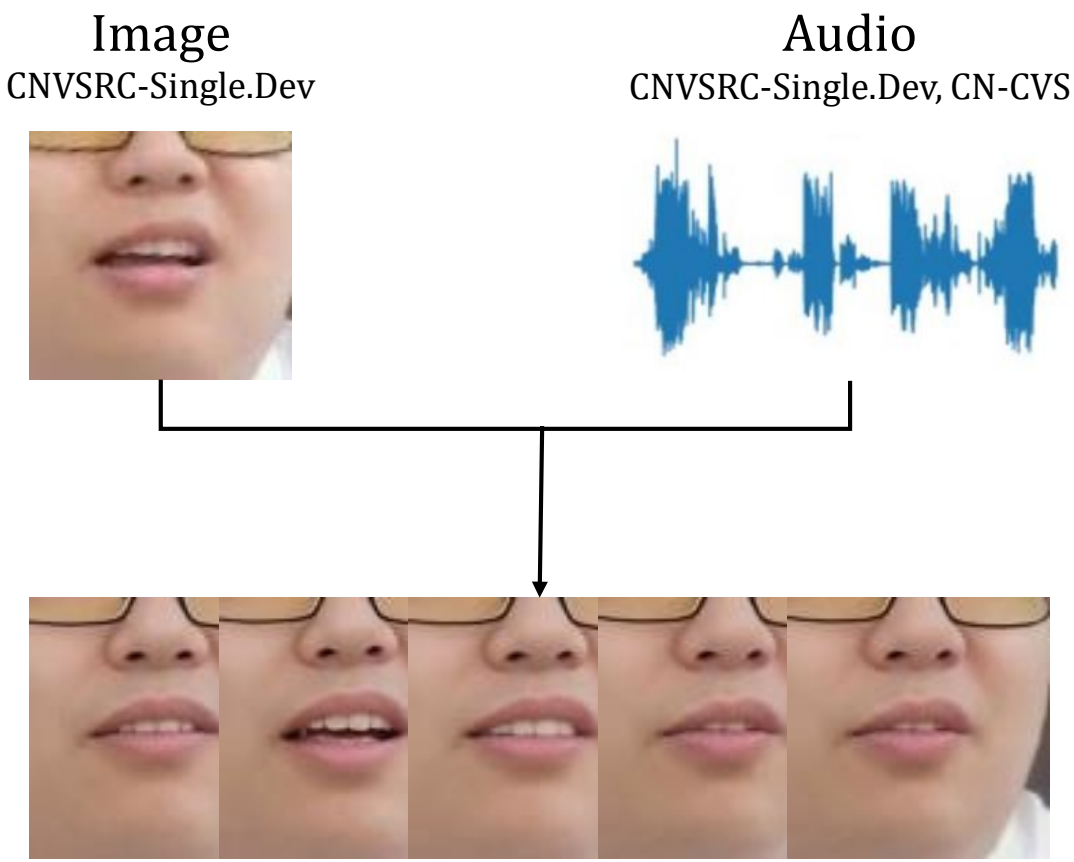
	Data Aug		Phoneme	LM	CER %
	ST	SP			
Base	✗	✗	✗	✗	48.60
S1	✓	✗	✓	✗	41.62
S2	✓	✗	✓	✓	40.64
S3	✓	✓	✓	✓	40.72
S4	Fusion of S2 and S3				40.51

Outline

- Data & Data Augmentation
  - SadTalker based data augmentation
  - Speed Perturbation based data augmentation
  - Data Usage
- Methods
  - Character-Phoneme Multi-Task Training Framework
  - Transformer-based Language Model
  - System Fusion
- Training Strategy
- Result and Conclusion
  - Overall T238 VII-BUPT System
  - Result

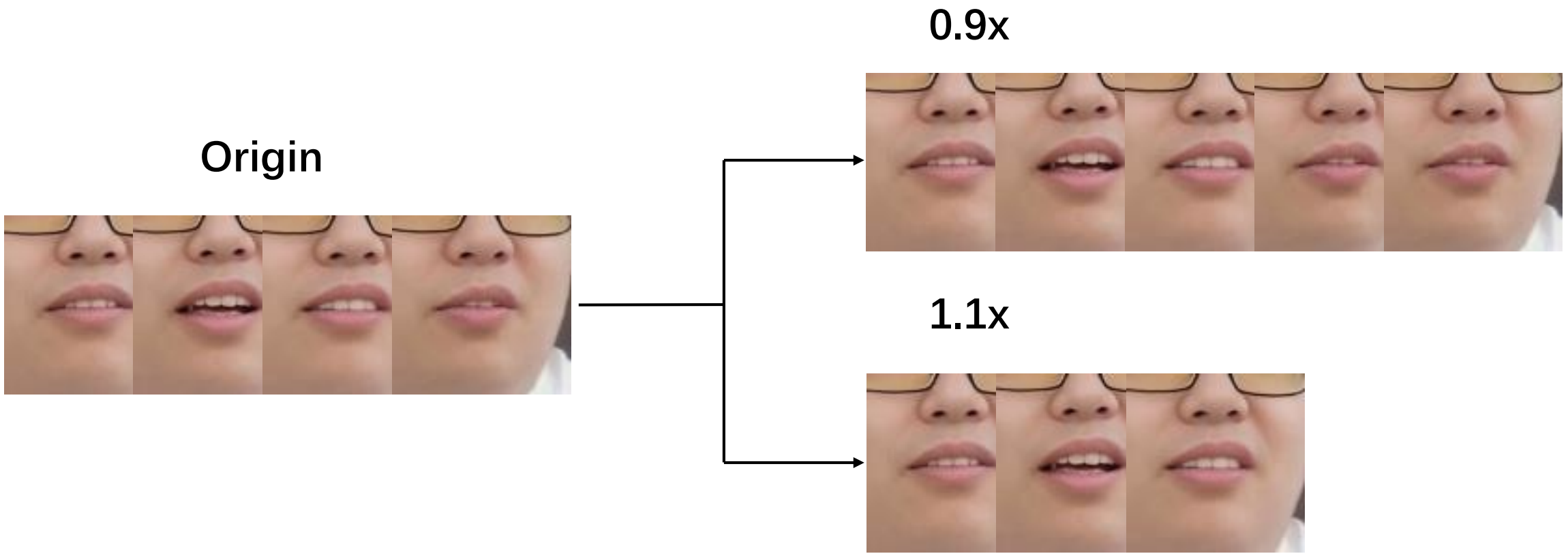
SadTalker-based Data Augmentation

- Audio driven lip generation

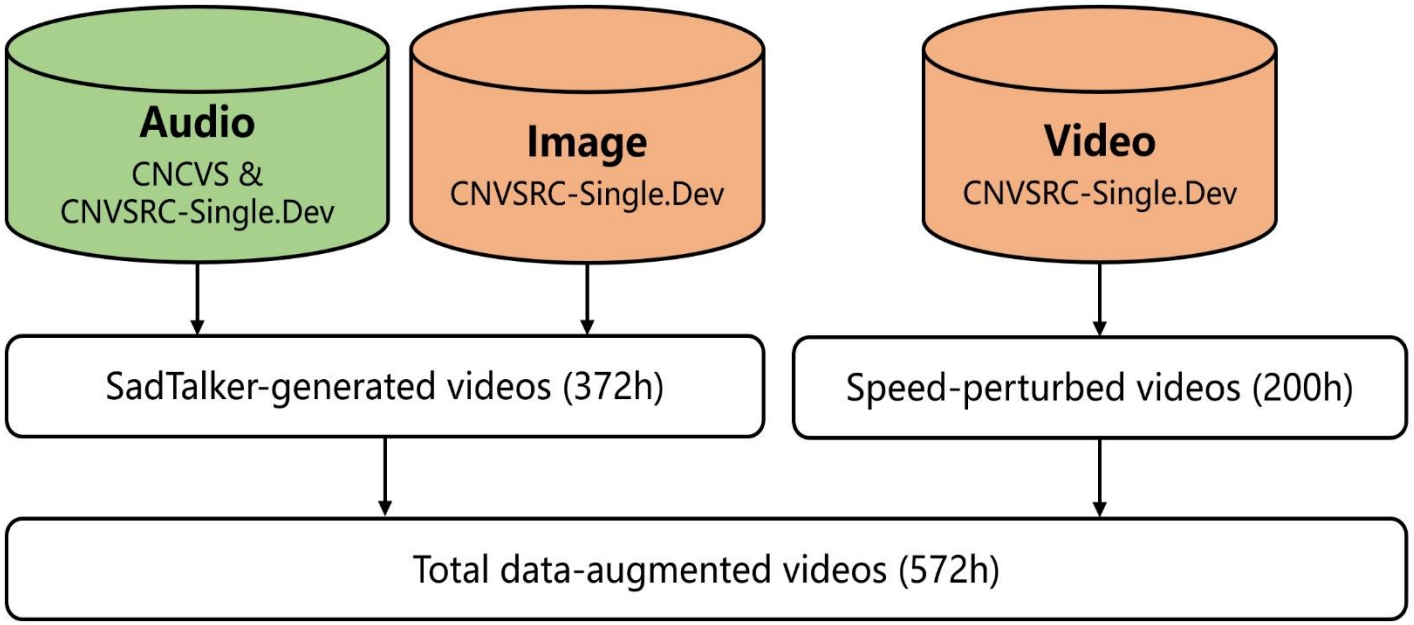


Speed Perturbation based Data Augmentation

- CNVSRC-Single.Dev
- Altering speeds to 0.9x and 1.1x



Data Usage

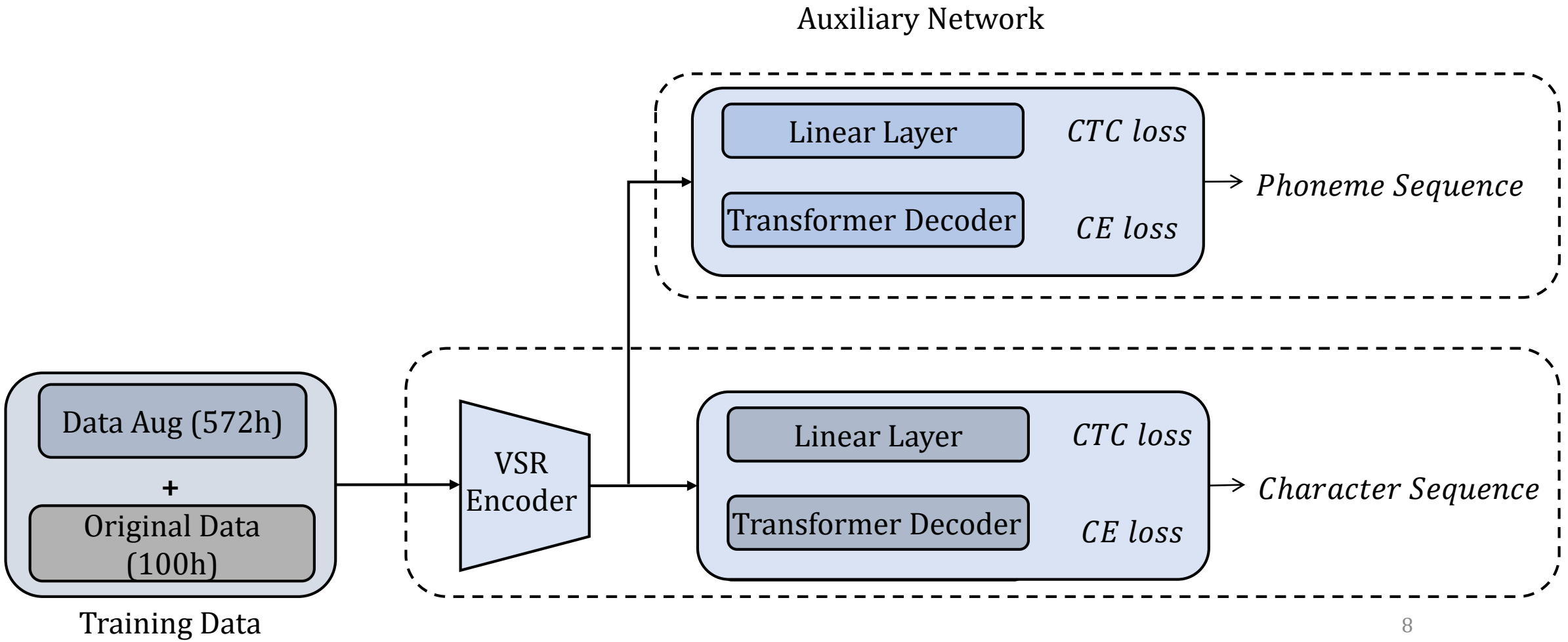

Audio
CNCVS &
CNVSRC-Single.Dev
Image
CNVSRC-Single.Dev
Video
CNVSRC-Single.Dev
SadTalker-generated videos (372h)
Speed-perturbed videos (200h)
Total data-augmented videos (572h)

Outline

- Data & Data Augmentation
  - SadTalker based data augmentation
  - Speed Perturbation based data augmentation
  - Data Usage
- **Methods**
  - Character-Phoneme Multi-Task Training Framework
  - Transformer-based Language Model
  - System Fusion
- Training Strategy
- Result and Conclusion
  - Overall T238 VII-BUPT System
  - Result

Character-Phoneme Multi-Task Training

- Considering the close correlation between pronunciation and lip shape
- $L = \mathrm{a}\left(mL_{\mathrm{p}-ctc} + (1-m)\,L_{p-att}\right) + b(nL_{c-ctc} + (1-n)L_{c-att})$



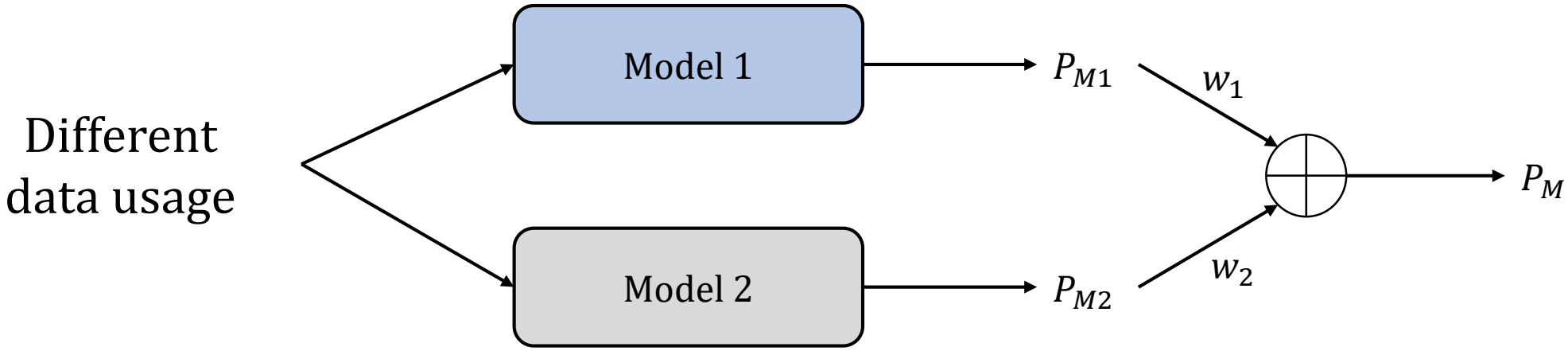
Transformer based Language Model

- Structure
  - Transformer-based
  - 8 attention heads, 16 layers.
- Data
  - CN-CVS (183,360) + CNVSRC-Single.Dev (25,946)
- Performance Improvement
  - CER improved from 41.62% to 40.64%.

	Data Aug		Phoneme	LM	CER %
	ST	SP			
Base	✗	✗	✗	✗	48.60
S1	✓	✗	✓	✗	41.62
S2	✓	✗	✓	✓	40.64
S3	✓	✓	✓	✓	40.72
S4	Fusion of S2 and S3				40.51

System Fusion

- Different models possess unique strengths and weaknesses
- System fusion can leverage the benefits of collective decision-making



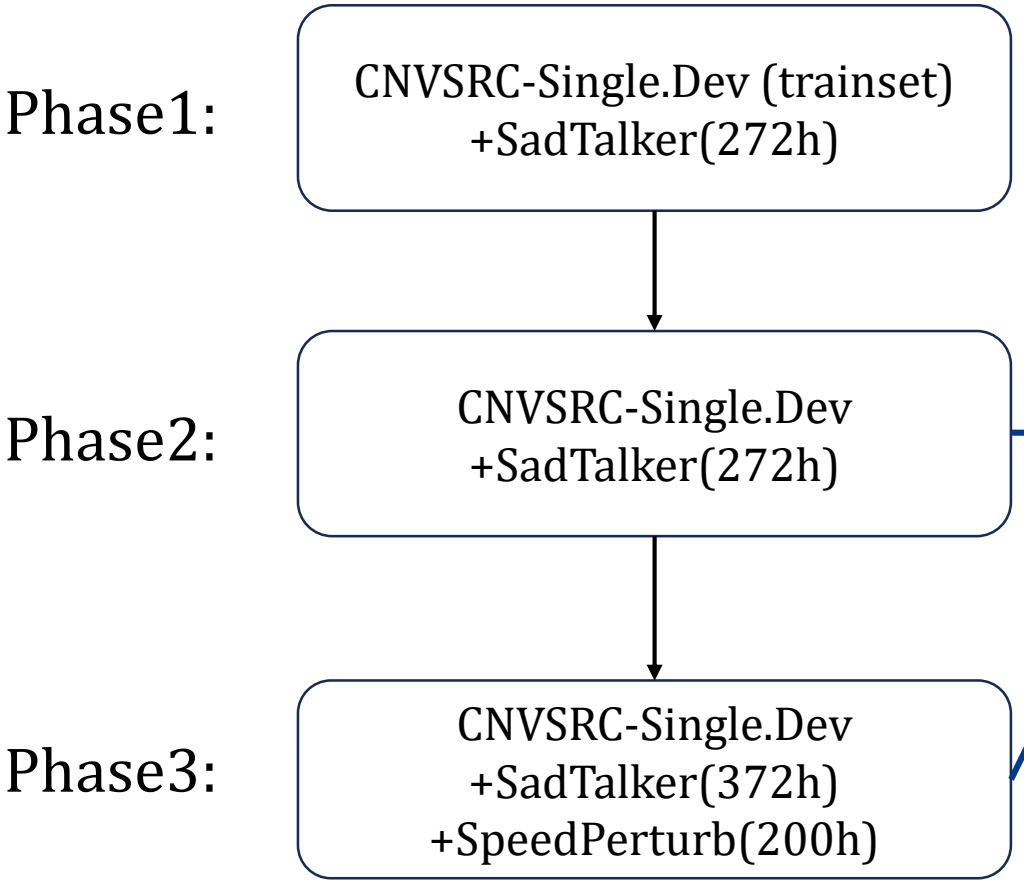
	Data Aug		Phoneme	LM	CER %
	ST	SP			
Base	✗	✗	✗	✗	48.60
S1	✓	✗	✓	✗	41.62
S2	✓	✗	✓	✓	40.64
S3	✓	✓	✓	✓	40.72
S4	Fusion of S2 and S3				40.51

Outline

- Data & Data Augmentation
  - SadTalker based data augmentation
  - Speed Perturbation based data augmentation
  - Data Usage
- Methods
  - Character-Phoneme Multi-Task Training Framework
  - Transformer-based Language Model
  - System Fusion
- **Training Strategy**
- Result and Conclusion
  - Final Framework
  - Result

Multi-Task Training

- AdamW optimizer
- Warm-up + Cosine Annealing learning rate scheduler



	Data Aug		Phoneme	LM	CER %
	ST	SP			
Base	✗	✗	✗	✗	48.60
S1	✓	✗	✓	✗	41.62
S2	✓	✗	✓	✓	40.64
S3	✓	✓	✓	✓	40.72
S4	Fusion of S2 and S3				40.51

Outline

- Data & Data Augmentation
  - SadTalker based data augmentation
  - Speed Perturbation based data augmentation
  - Data Usage
- Methods
  - Character-Phoneme Multi-Task Training Framework
  - Transformer-based Language Model
  - System Fusion
- Training Strategy
- **Result and Conclusion**
  - Overall T238 VII-BUPT System
  - Result

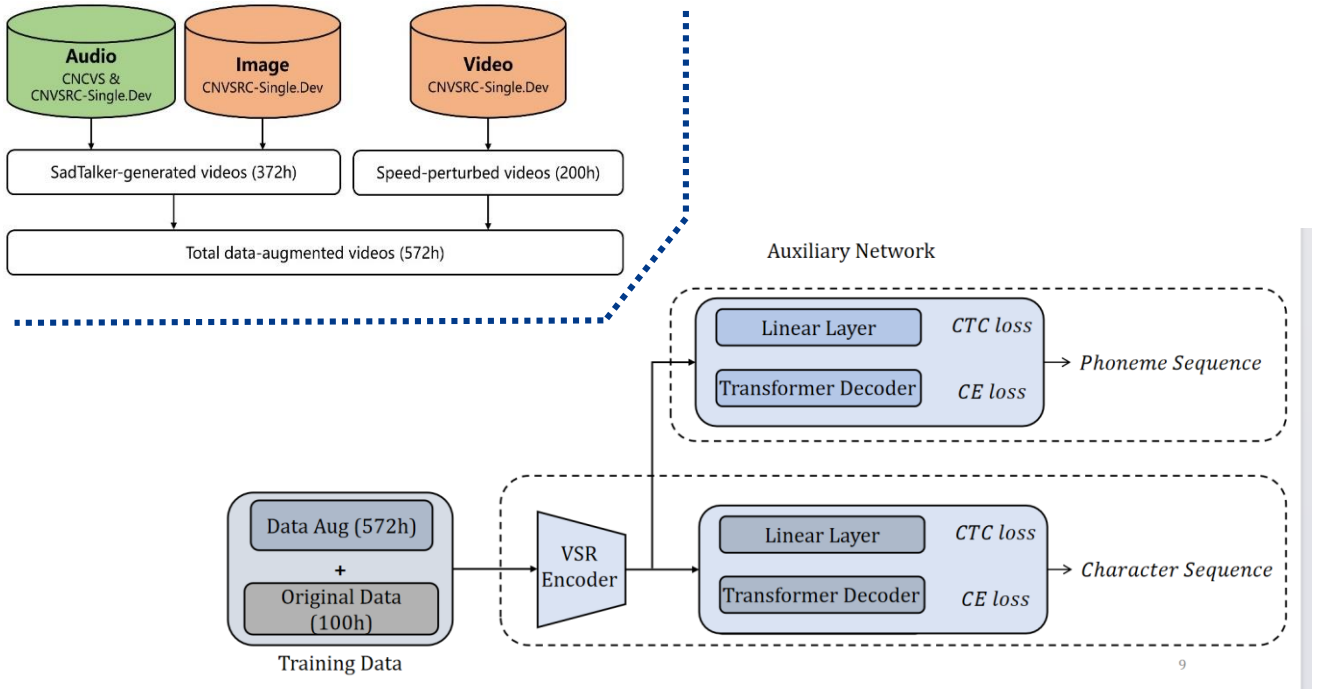
Overall T238 VII-BUPT System

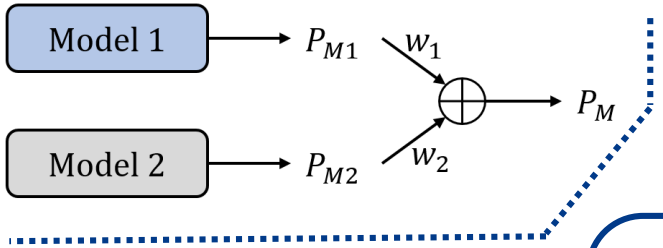
Data Augmentation

Character-Phoneme VSR Model

System Fusion

Language Model





	Data Aug		Phoneme	LM	CER %
	ST	SP			
Base	✗	✗	✗	✗	48.60
S1	✓	✗	✓	✗	41.62
S2	✓	✗	✓	✓	40.64
S3	✓	✓	✓	✓	40.72
S4	Fusion of S2 and S3				40.51

Result

- **ST**: SadTalker-based data augmentation
- **SP**: Speed Perturbation data augmentation
- **LM**: Language Model during prediction

	Data Aug		Phoneme	LM	CER %
	ST	SP			
Base	✗	✗	✗	✗	48.60
S1	✓	✗	✓	✗	41.62
S2	✓	✗	✓	✓	40.64
S3	✓	✓	✓	✓	40.72
S4	Fusion of S2 and S3				40.51

Thanks