Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert

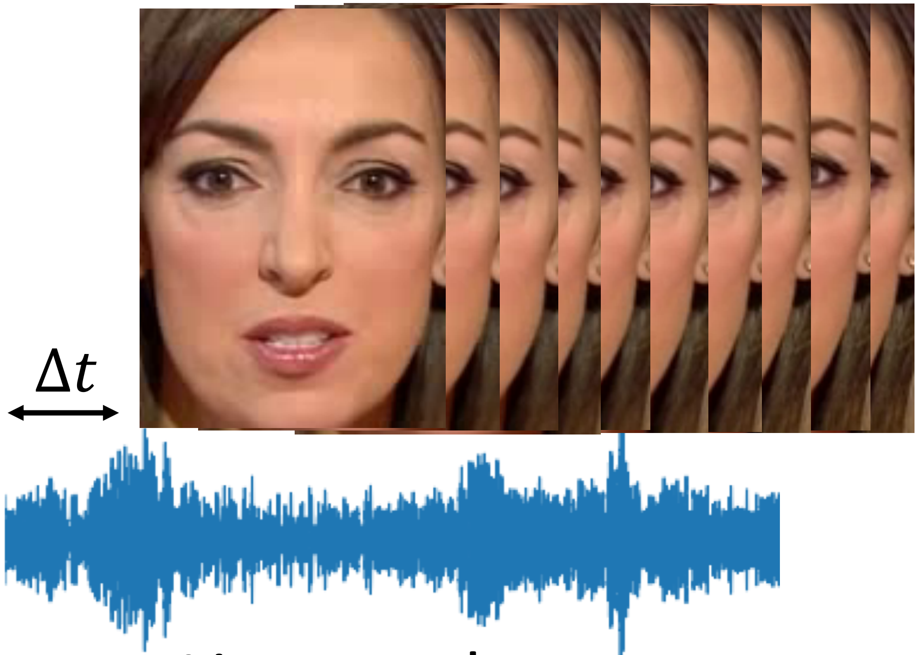
Jiadong Wang, Xinyuan Qian, Malu Zhang,
Robby T. Tan, Haizhou Li

Paper ID: 10744
Session: WED-PM-220

NUS
National University
of Singapore

Introduction

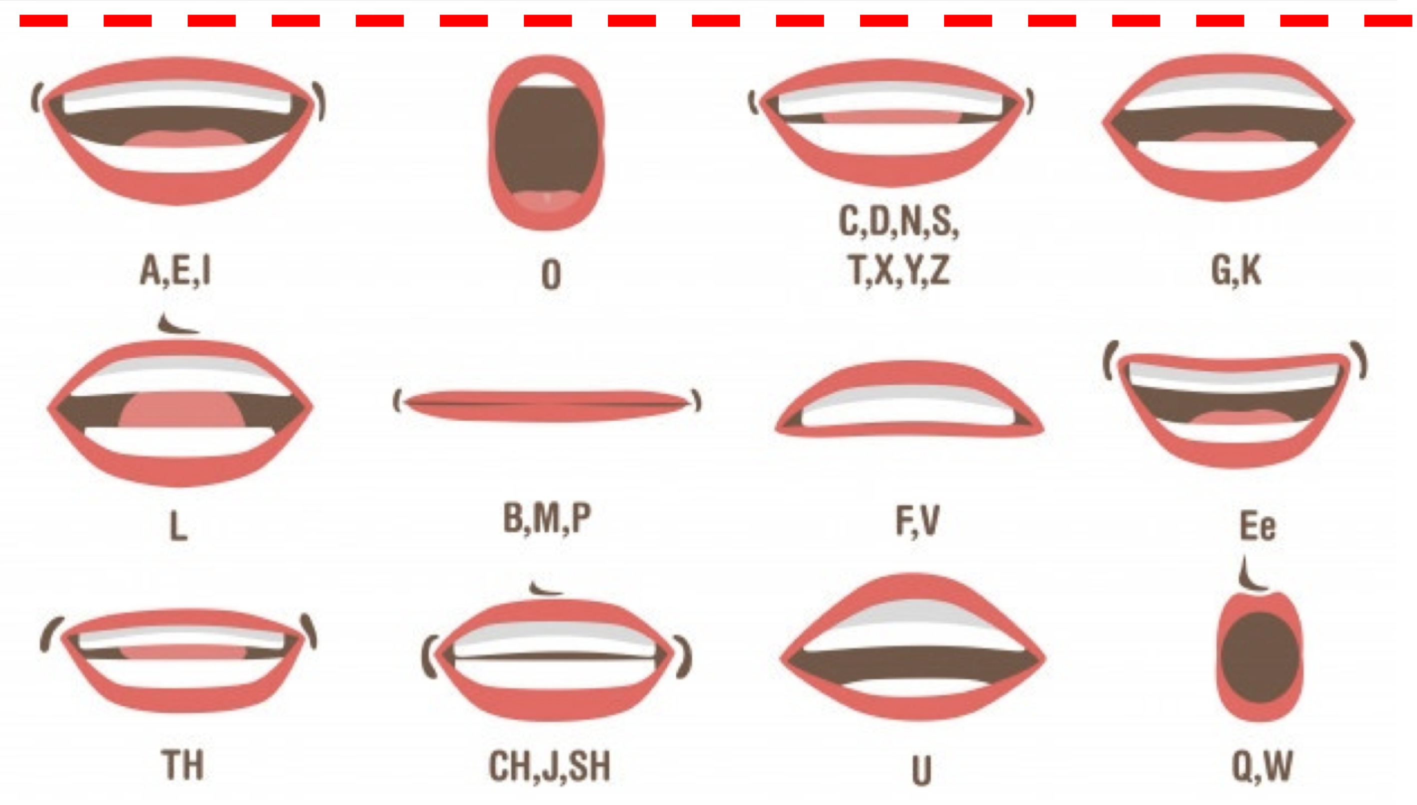
Visual quality



Lip-speech sync

- Visual quality and lip-speech sync are widely concerned aspects of talking face generation.

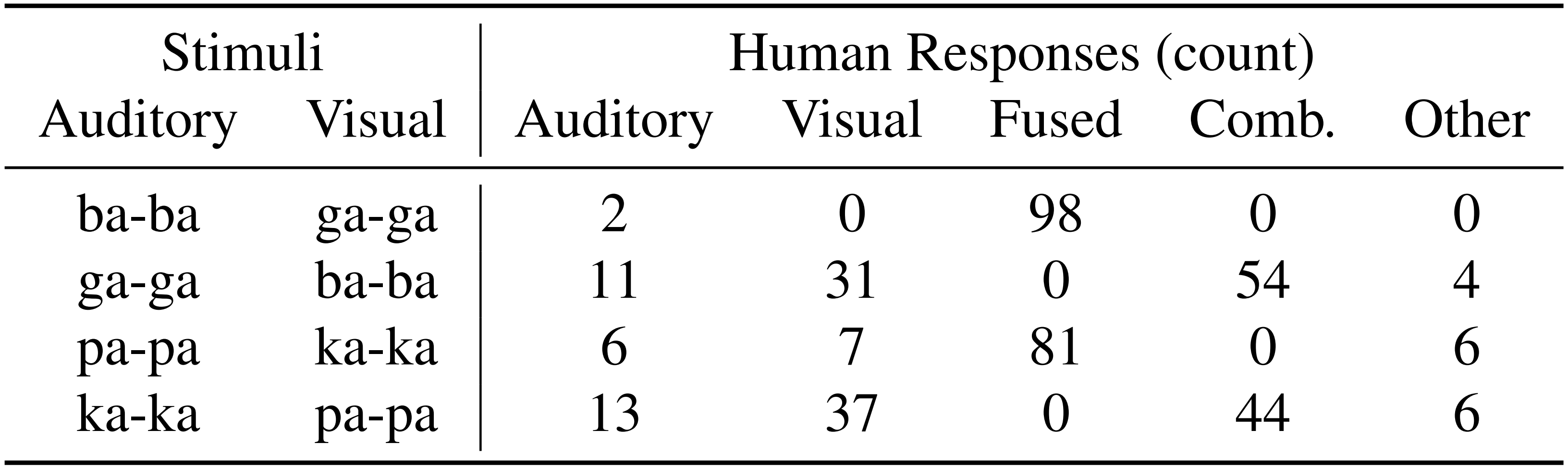
Introduction



Reading Intelligibility

- Reading intelligibility indicates how much text content can be interpreted from lip movements.

Introduction



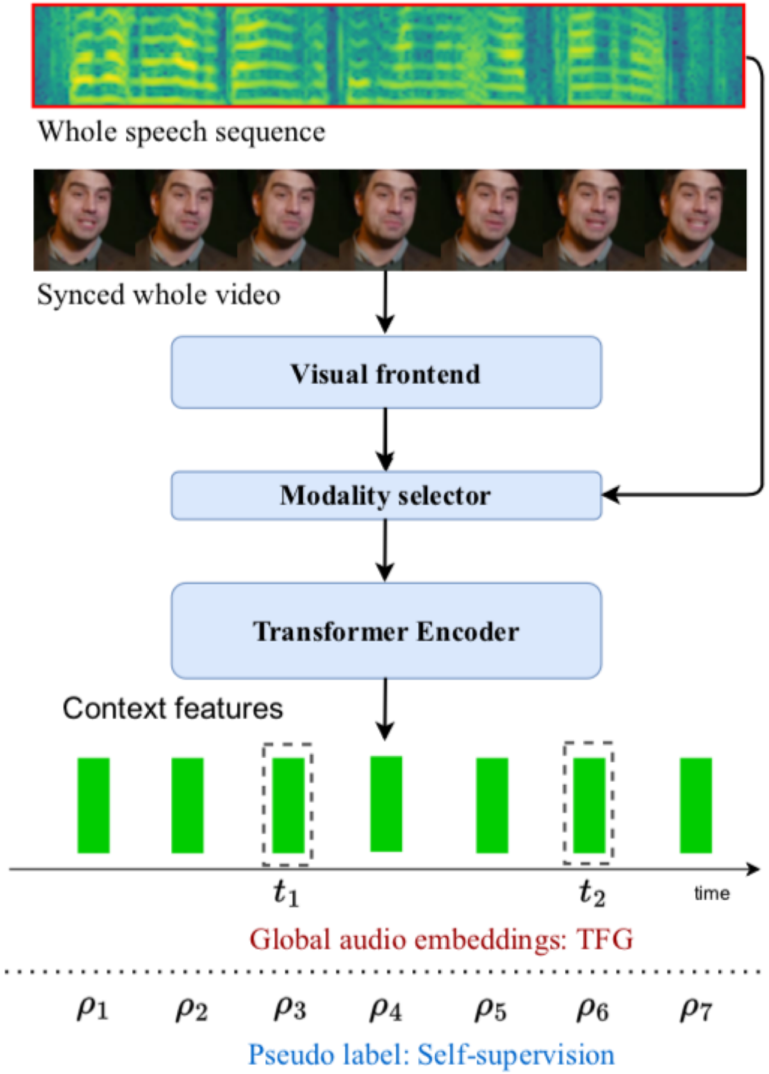
Stimuli		Human Responses (count)				
Auditory	Visual	Auditory	Visual	Fused	Comb.	Other
ba-ba	ga-ga	2	0	98	0	0
ga-ga	ba-ba	11	31	0	54	4
pa-pa	ka-ka	6	7	81	0	6
ka-ka	pa-pa	13	37	0	44	6

McGurk Effect

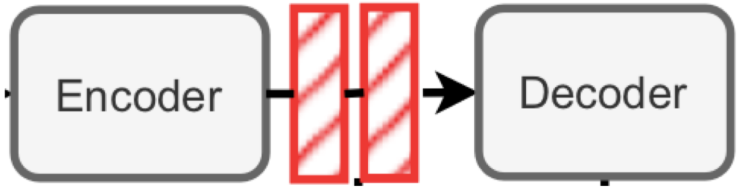
- Visual quality and lip-speech synchronization do not explicitly reflect intelligibility.

Overview of lip-reading expert

Self-supervised Pre-training


Whole speech sequence
Synced whole video
Visual frontend
Modality selector
Transformer Encoder
Context features
t_1
t_2
time
Global audio embeddings: TFG
ρ_1 ρ_2 ρ_3 ρ_4 ρ_5 ρ_6 ρ_7
Pseudo label: Self-supervision

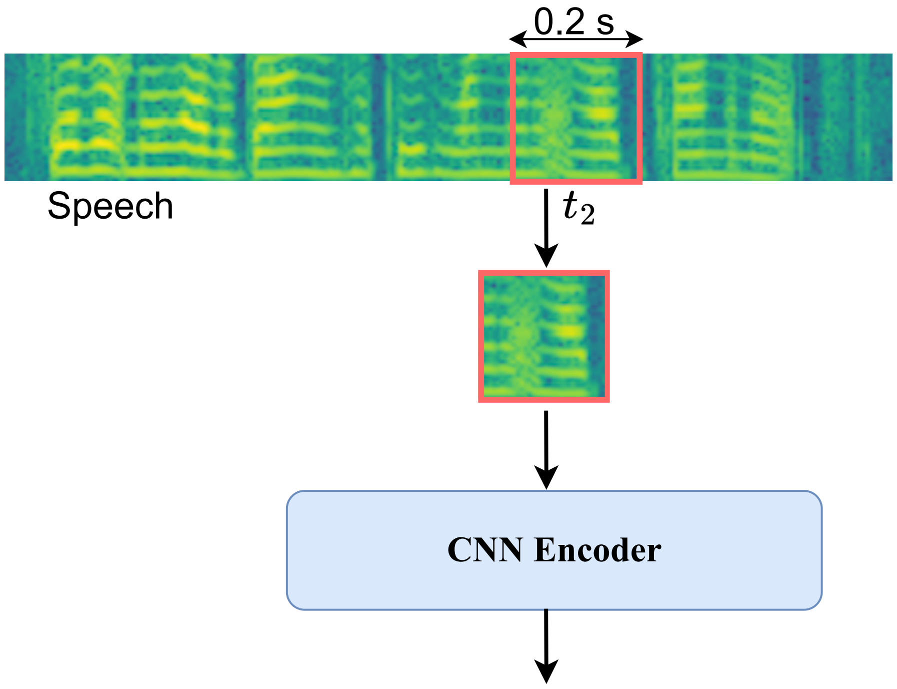
Supervised Fine-tuning



- **Self-supervised Pre-training** uses the clustering class of hand-crafted audio feature or learned audio-visual feature as **pseudo labels**.

- **Supervised Finetuning** constructs a lip-reading experts with **the pre-trained transformer encoder** and **a decoder** and trains it with **text annotation**.

Audio encoder

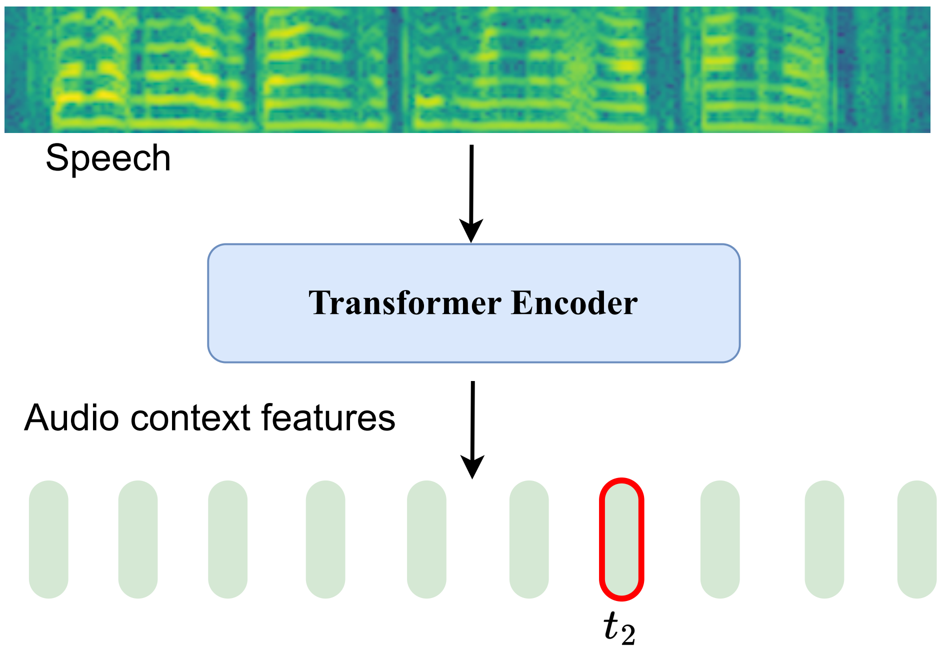


0.2 s

Speech

t_2

CNN Encoder



Speech

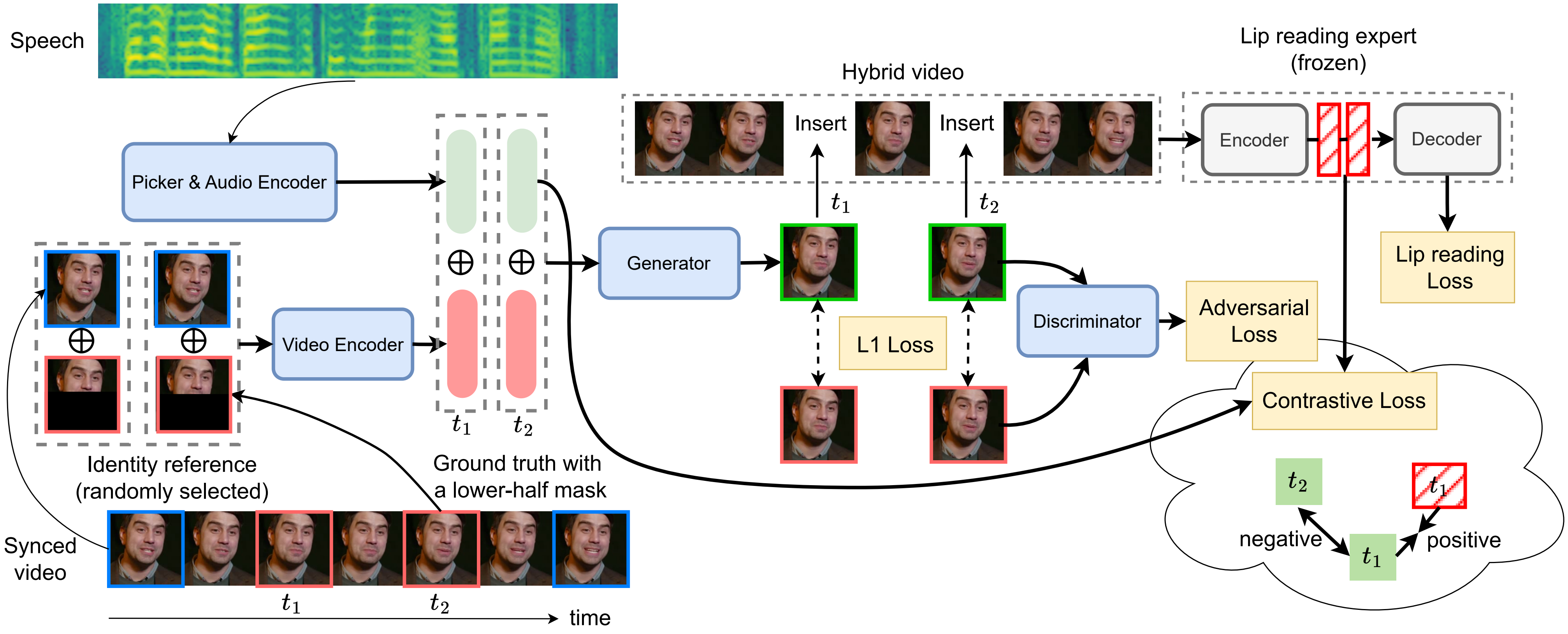
Transformer Encoder

Audio context features

t_2

- **Local audio embedding** crop a **0.2s** audio segment whose **centre** is **temporally aligned** with **an input image**.
- **Global audio embedding** extract audio context features from an **entire audio** and then crop a feature which is **temporally aligned** with **an input image**.

Architecture



- Synthesis of talking face given a triplet of **a pose image, an identity image** and **a speech.**
- Penalize incorrect lip movements in synthesized image via a lip reading expert.
- **Contrastive learning** between **audio embeddings** and output **features** of the **lip reading expert's encoder**.

Contribution

- We tackle the **reading intelligibility** problem of speech-driven talking face generation by **leveraging a lip-reading expert**.

- To **enhance lip-speech synchronization**, we propose a novel cross-modal **contrastive learning** strategy, **assisted by a lip-reading expert**.

- We employ **a transformer encoder** trained **synchronically** with the **lip-reading expert** to consider **global temporal dependency** across the entire audio utterance.

- We propose a new strategy to **evaluate reading intelligibility** and **make the benchmark code publicly available**.

- Extensive experiments show that our proposal achieve **SOTA reading intelligibility and lip-speech synchronization.**

Experiments

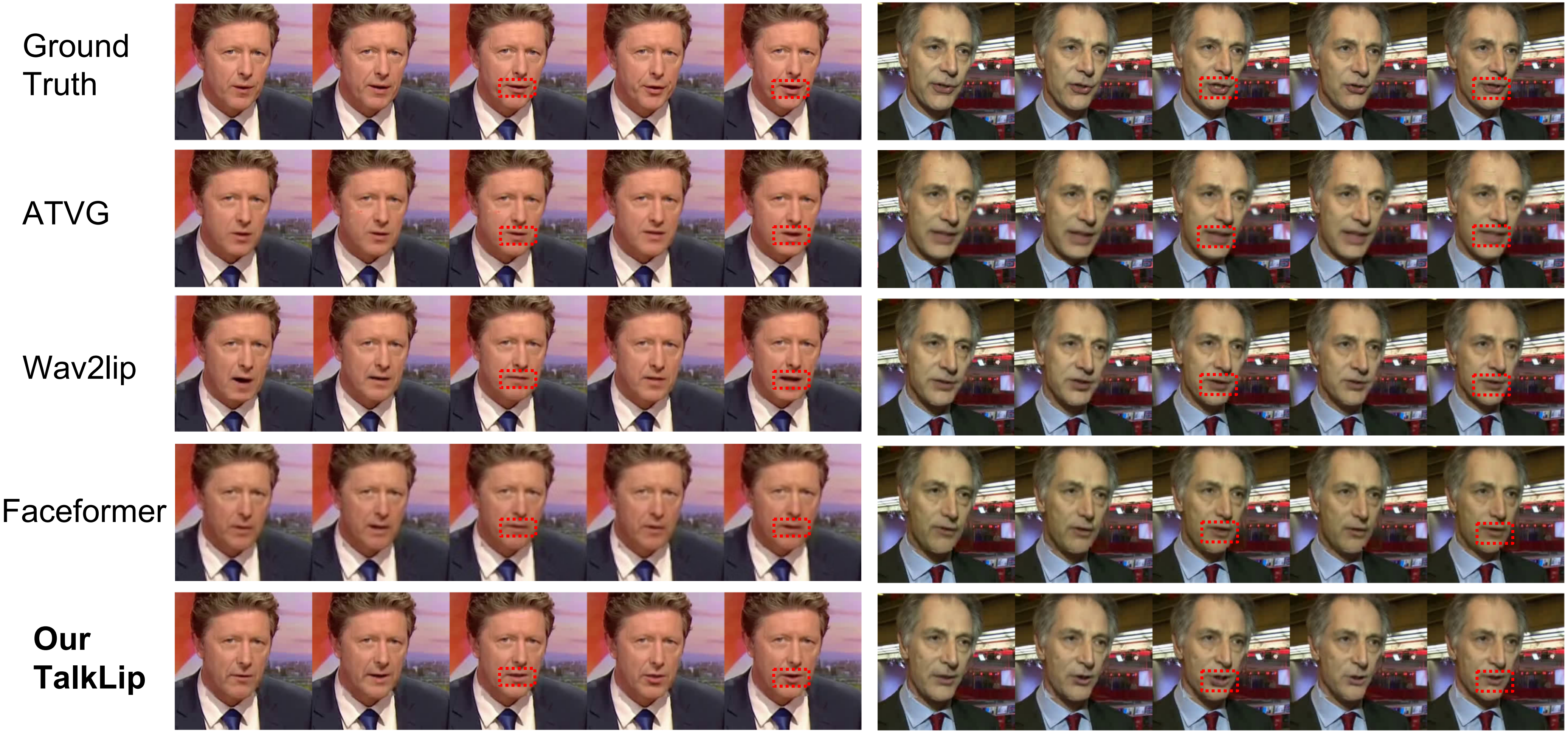
- **Training dataset**
  - LRS2 train set (29 hours)
- **Evaluation dataset**
  - LRS2 test set: continuous audio-visual speech recognition
  - LRW test set: audio-visual word classification
- **Metrics**
  - **Visual quality**:
    - SSIM
    - PSNR
  - **Lip-speech synchronization**:
    - LSE-C
    - LSE-D
  - **Reading intelligibility**:
    - Word Error Rate on **LRS2**
    - Accuracy on **LRW**

Quantitative Result

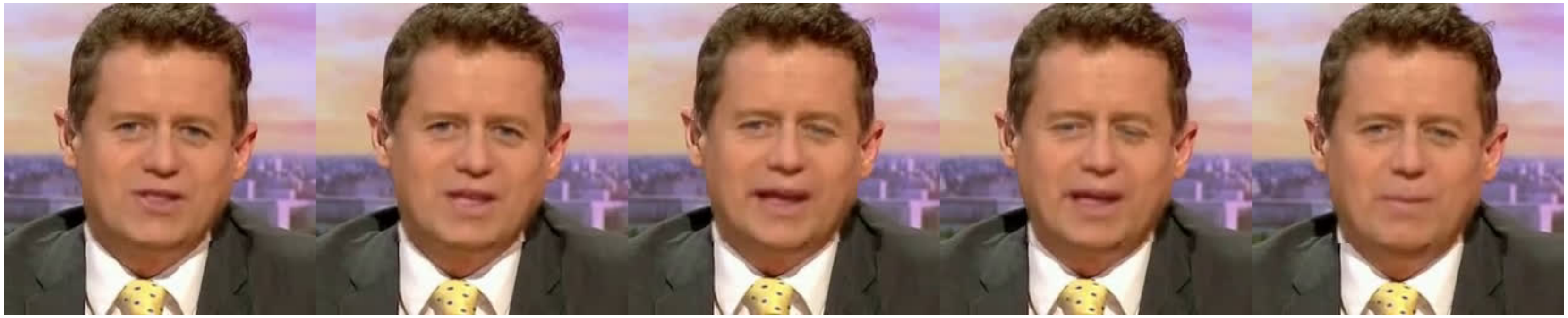
Method	LRW				LRS2				
	PSNR ↑	SSIM ↑	LSE-C ↑	ACC (%) ↑	PSNR ↑	SSIM ↑	LSE-C ↑	WER$_1$ (%)	WER$_2$ ↓
Ground Truth	N.A.	1.000	6.88	88.51	N.A.	1.000	8.25	23.82	40.9
ATVGnet	30.71	0.791	5.64	18.10	30.42	0.751	5.05	113.69	91.8
Wav2Lip	31.52	0.874	7.18	59.98	31.36	0.854	8.40	82.06	73.9
Faceformer	29.19	0.856	5.58	53.43	29.47	0.840	6.42	97.64	79.0
PC-AVS*	30.44	0.778	6.42	-	29.89	0.747	6.73	-	-
SyncTalkFace*	**33.13**	**0.893**	6.62	-	**32.59**	**0.876**	7.93	-	-
TalkLip (l)	31.24	0.867	6.44	79.78	31.38	0.849	7.58	45.74	55.7
TalkLip ($l + c$)	31.52	0.867	6.51	83.17	31.14	0.850	7.76	38.00	49.2
TalkLip (g)	30.78	0.871	7.01	86.57	30.86	0.854	8.38	25.31	36.5
TalkLip ($g + c$)	31.18	0.866	**7.28**	**87.81**	31.19	0.850	**8.53**	**23.43**	**35.1**
Base w.o.$\mathcal{L}_{\text{lip}}$	31.22	0.865	6.01	48.58	31.08	0.852	7.09	103.57	82.2
Base w.o.$\mathcal{L}_{\text{lip,gan}}$	30.64	0.864	5.03	30.80	30.70	0.851	5.93	116.26	89.3

- g and l: global and local audio embedding
- c: Contrastive learning
- **Base** denotes **Talklip** (l)
- * indicates that results are scratched from another paper as these methods do not open-source their training scripts.

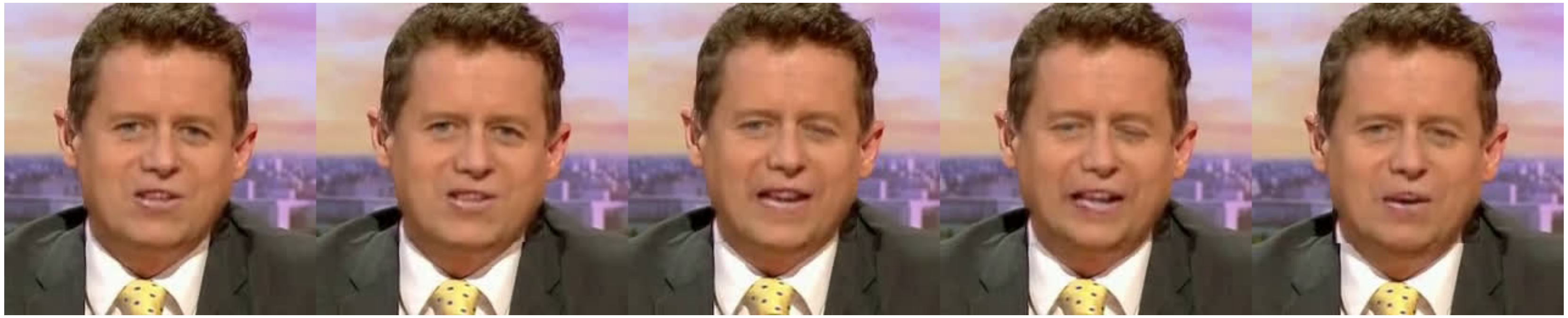
Qualitative Result



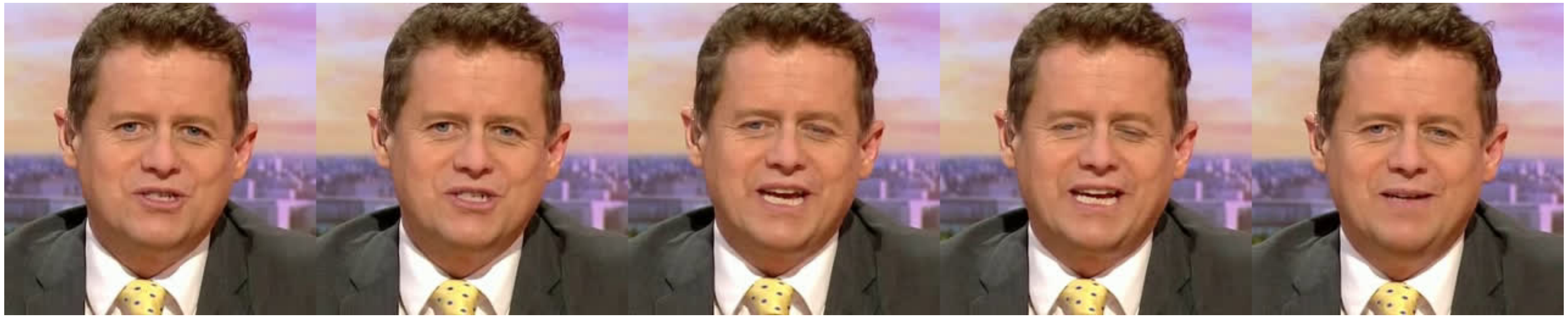
Ablation on Audio Encoder



a) TalkLip (*l*+*c*)

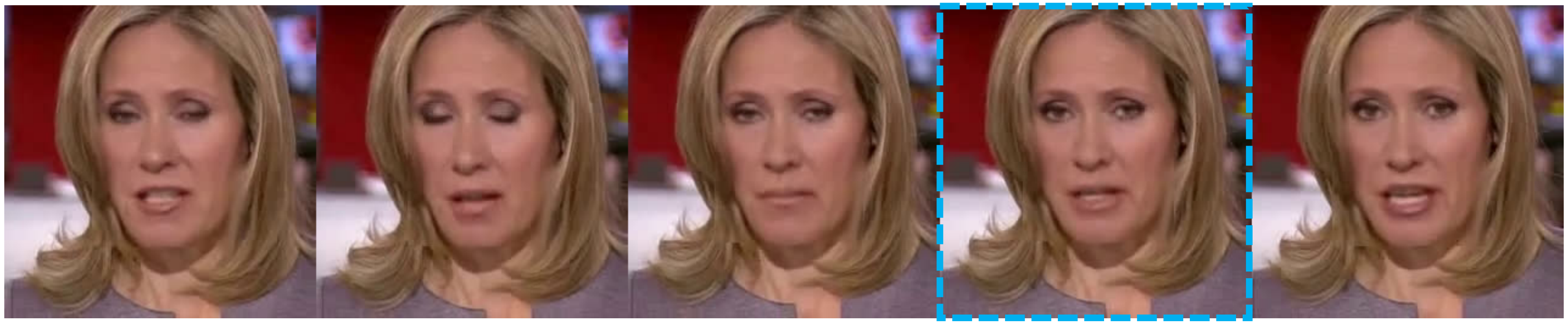


b) TalkLip (*g*+*c*)

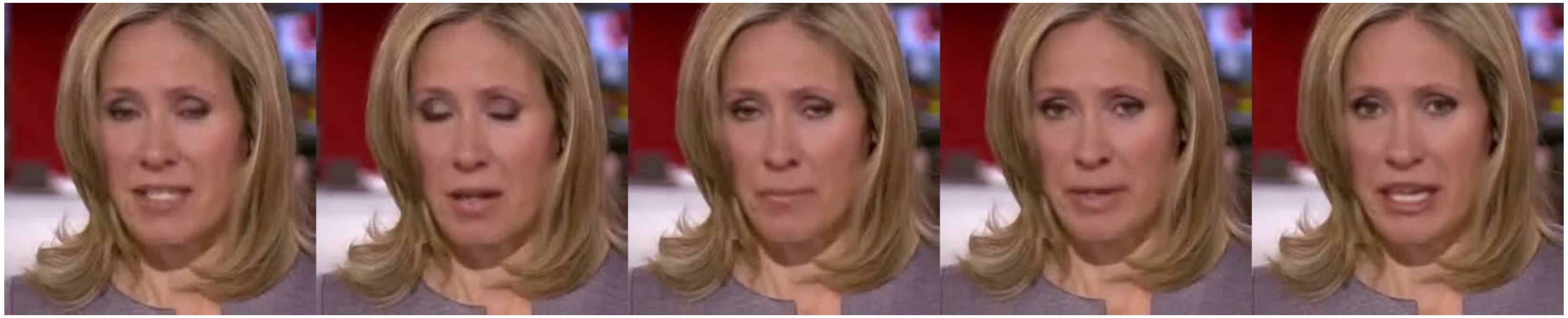


c) Ground Truth

Ablation on Contrastive Learning



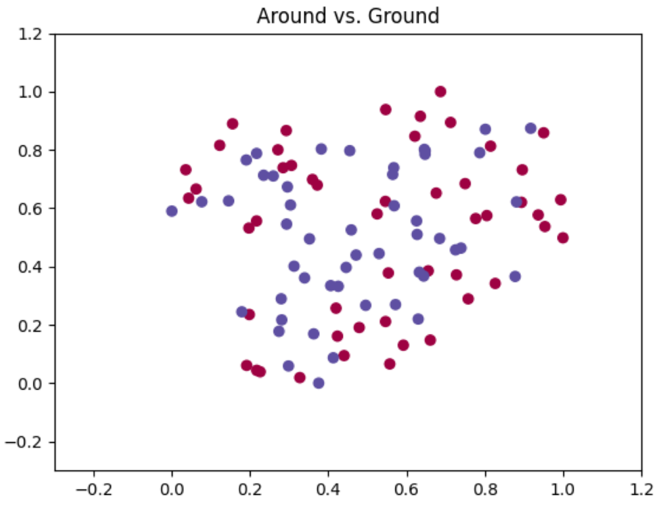
a) TalkLip (*l*)



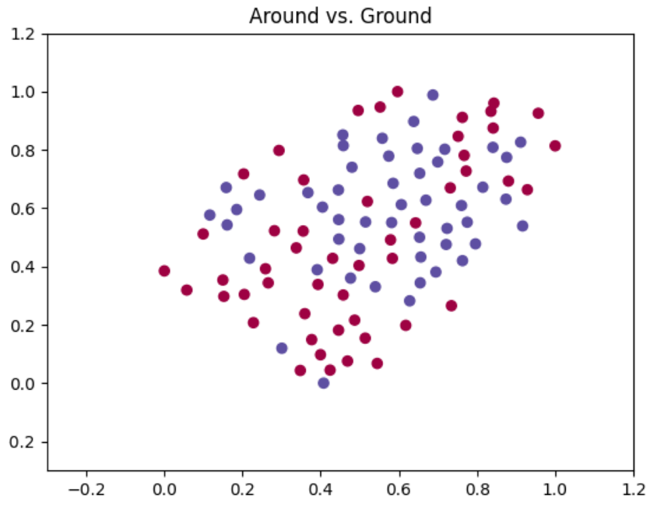
b) TalkLip (*l*+*c*)

c) Ground Truth

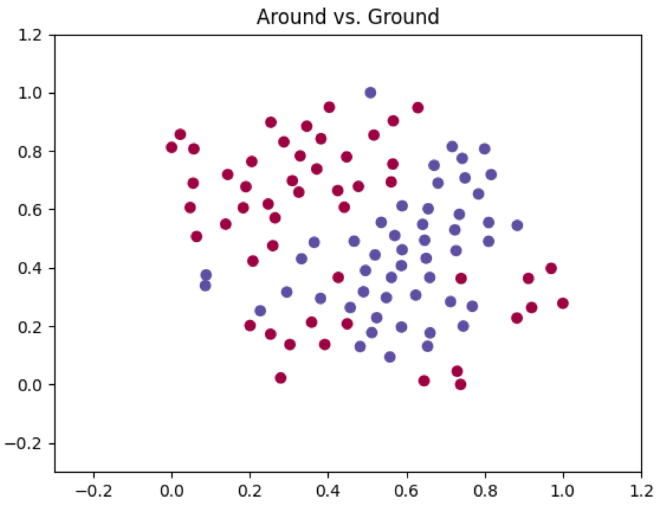
Audio Embedding Visualization



Talklip (l)



Talklip ($l + c$)



Talklip ($g + c$)

Demo

What’s the Best Thing about the Royal Highland show

Conclusion

- A lip reading expert is efficient to improve reading intelligibility.
- The contrastive learning can boost not only lip-speech synchronization but also reading intelligibility.
- The transformer encoder can both improve reading intelligibility and lip-speech synchronization.
- Extensive experiments prove that our proposal achieve **SOTA reading intelligibility and lip-speech synchronization.**