THU-AM-222

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

SynthVSR: Scaling Up Visual Speech Recognition With Synthetic Supervision

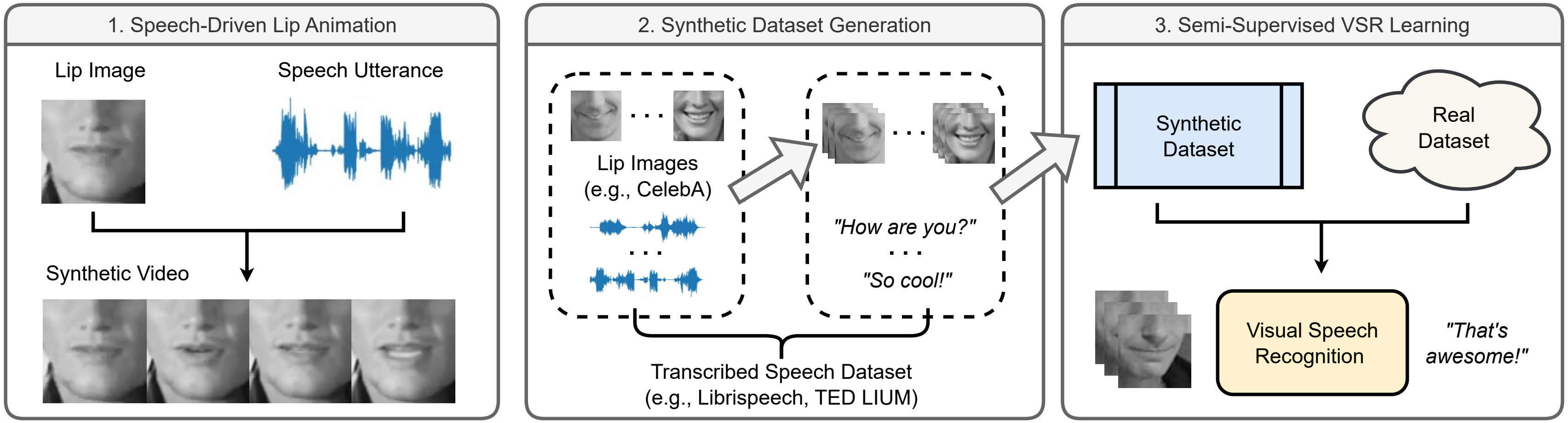
Xubo Liu[1,2], Egor Lakomkin[2], Konstantinos Vougioukas[2], Pingchuan Ma[2], Honglie Chen[2], Ruiming Xie[2], Morrie Doulaty[2], Niko Moritz[2], Jáchym Kolář[2], Stavros Petridis[2], Maja Pantic[2], Christian Fuegen[2]

[1]University of Surrey [2]Meta AI

1-Minute Summary

- Key challenge of visual speech recognition (VSR)
  - Lack of large-scale labeled audio-visual video data (e.g., LRS3 438 hours)

- SynthVSR: Proposed semi-supervised framework for VSR
  - Generate synthetic lip movement videos from speech and face datasets
  - SOTA WER 16.9% is achieved on LRS3, using 29x less data than previous methods



Related Work

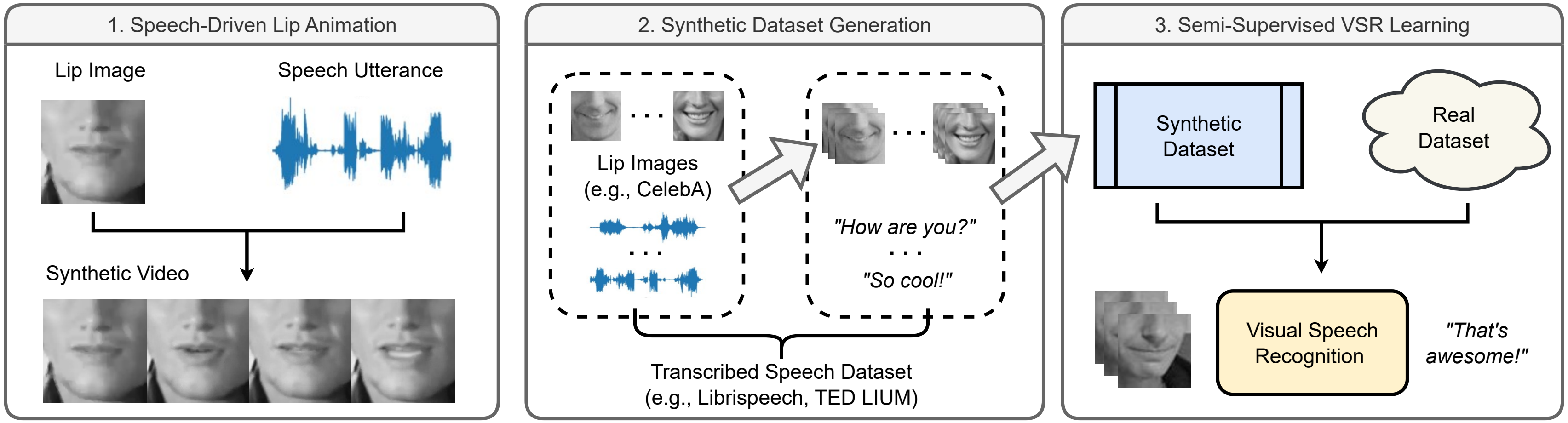
- Recently SOTA methods - **leverage increasingly large amount of audio-visual data**
  - Supervised learning
    - Collect large-scale non-public transcribed audio-visual datasets e.g., 90,000 hours of data from Google (Serdyuk et al. 2022)
  - Semi-supervised learning
    - Use ASR to label audio-visual data (Ma et al. 2022):
  - Self-supervised learning - AV-HuBERT (Shi et al. 2022)
    - Self-supervised learning on 1700 hours of unlabelled audio-visual data

Related Work

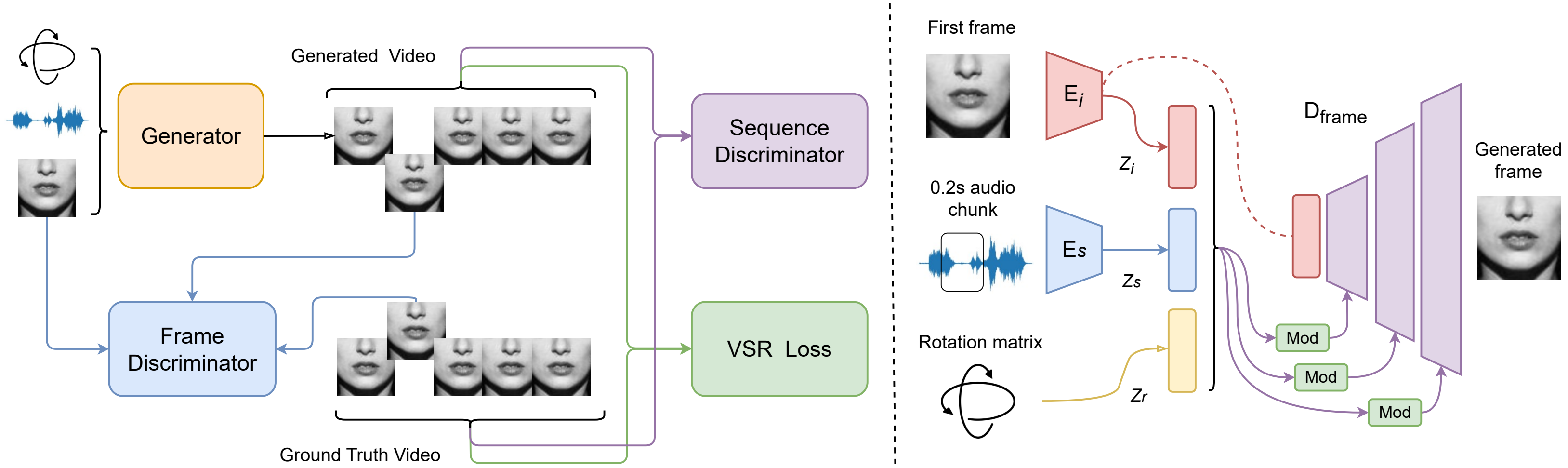
- Publicly available audio-visual dataset is limited in size
  - Limitations:
    - Includes speaker’s frontal face
    - Resource-intensive to collect, not easy to scale up
    - Privacy and bias issues
    - Licensing for industry research (e.g., LRS2, LRW)

SynthVSR: Scaling Up VSR With Synthetic Supervision

- **SynthVSR (Proposed semi-supervised VSR framework)**
  - Speech-driven lip animation
    - Generate synthetic video clip from speech signal and face image
  - Practically infinite data diversity (identities, text labels) for scaling up VSR



Speech-Driven Lip Animation (VSR-Oriented)



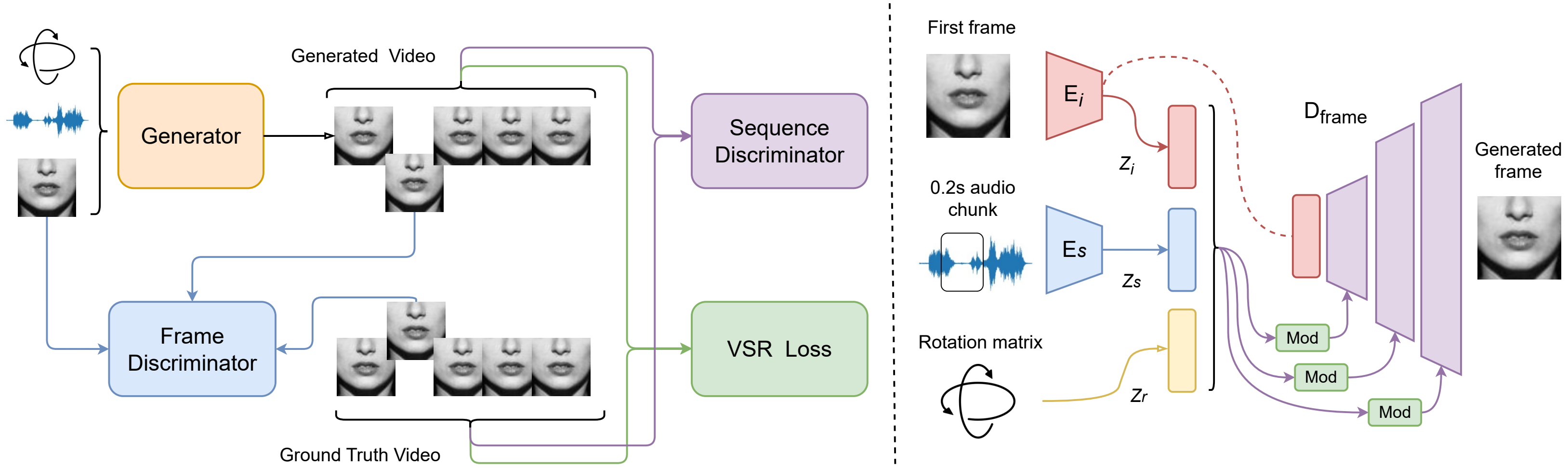
- Temporal GAN with two discriminators (frame & sequence)

$$\mathcal{L}_{Disc}^{img} = \mathbb{E}_v[\log D_{img}(S(v), v_1)] + \mathbb{E}_{v,s}[\log(1 - D_{img}(S(G(s, v_1)), v_1)] \quad (1)$$

$$\mathcal{L}_{Disc}^{seq} = \mathbb{E}_v[\log D_{seq}(v)] + \mathbb{E}_{v,s}[\log(1 - D_{seq}(G(s, v_1))] \quad (2)$$

** **v**: video, **v_1**: first frame, **S(v)**: sampling function, **D**: Discriminator, **G**: Generator*

Speech-Driven Lip Animation (VSR-Oriented)

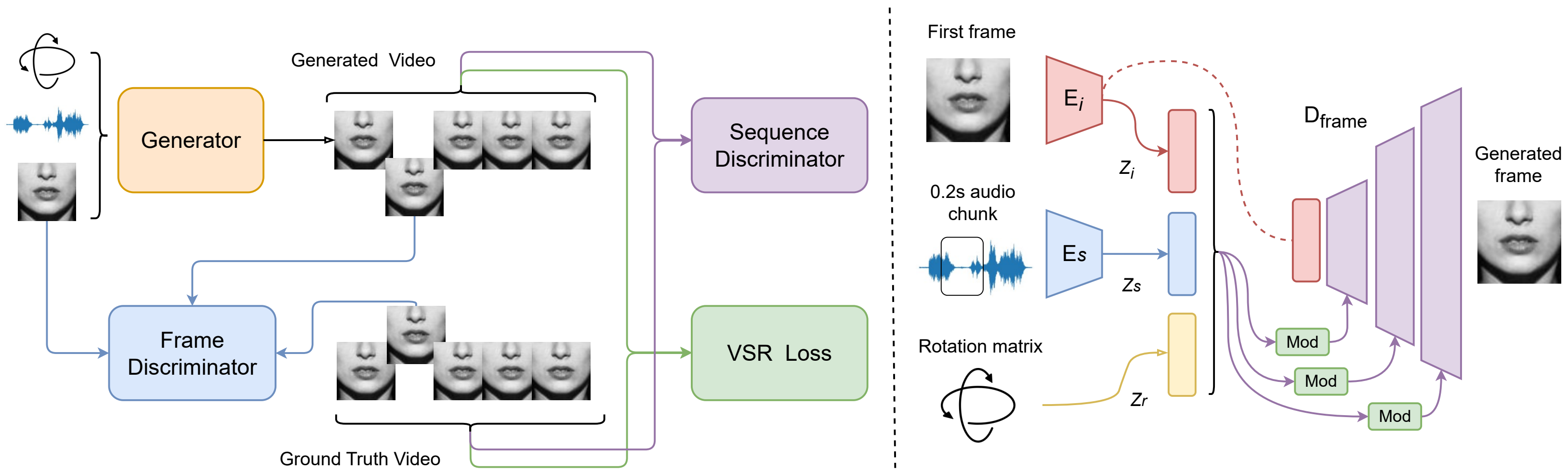


- A VSR perceptual loss is proposed when labelled videos are available:
  - Visual embedding L1 loss + linguistic logits KL loss

$$\mathcal{L}_{VSR} = \lambda_{visual} \left\| z_f^r - z_f^s \right\|_1 + \lambda_{logits} \, \mathrm{KL}(\hat{y}^r, \hat{y}^s) \quad (3)$$

* ***z_f***: *VSR visual embedding,* ***y_head***: *VSR predicted logits,* ***r***: *for real data,* ***s***: *for synthetic data*

Speech-Driven Lip Animation (VSR-Oriented)



- Pixel-level reconstruction loss:

$$\mathcal{L}_{rec} = \|v - \hat{v}\|_1 \qquad (4)$$

- Training objective:

* ***v_head****: synthetic video*

$$\mathcal{L}_{Animation} = \min_{\text{Gen.}} \max_{\text{Disc.}} (\lambda_{disc}^{img} \mathcal{L}_{disc}^{img} + \lambda_{disc}^{seq} \mathcal{L}_{disc}^{seq}) + \lambda_{rec} \mathcal{L}_{rec} + \mathcal{L}_{VSR} \qquad (5)$$

Scalable VSR Data Generation Pipeline

- Data generation pipeline:
  - Speech corpora (3,652 hours)
    - Librispeech, TED-LIUM, Common Voice
  - Face source:
    - CelebA (10k identities)
  - Generate one synthetic video per speech clip with one lip image
- Pros:
  - Easy to scale up
  - Unlimited data generation pipeline

Examples of synthetic lip movement videos

Experimental Setups

- Training data of speech-driven lip animation:
  - LRS3 + AVSpeech
- VSR model (Conformer-Transformer)
  - BASE (250M), LARGE (783M)
- Training and evaluation configuration is consistent with Ma et al. 2022
  - Training objective: CTC + CE
  - External LM is used for evaluation

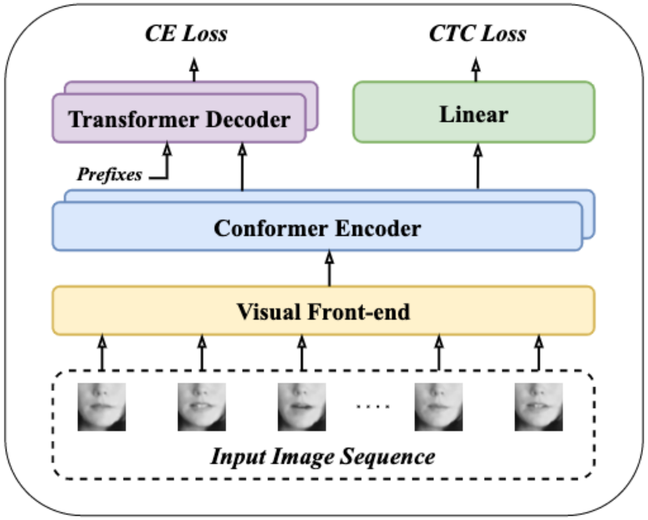


Figure 2. The VSR model used in this work based on a Conformer encoder, a 3D ResNet visual front-end and a combination of CTC and attention-based decoder.

Experimental Setups

- Benchmark : LRS3 - contains 408, 30, 0.9 hours of video clips from TED talks in the pre-training, training-validation and test set, respectively.
- Evaluation with multiple labelled data setups:
  - **Low-resource setup***:* 30 hours of LRS3
  - **Benchmark Setup***:* 438 hours of LRS3
  - **High-resource setup***:* 438 hours of LRS3 + 2630 hours of ASR pseudo-labelled public audio-visual data (Ma et al. 2022)

Meta AI

UNIVERSITY OF SURREY

Experimental Results – Low-Resource Setting

Method	Backbone	LM	Labeled data (hrs)	Unlabeled data (hrs)	Synthetic data (hrs)	WER (%)
Afouras et al. [3]	CNN	✓	595‡	334	-	59.8
Ren et al. [37]	Transformer	✗	818‡	-	-	59.0
Afouras et al. [1]	Transformer	✓	1,519†‡	-	-	58.9
Xu et al. [52]	RNN	✗	595‡	-	-	57.8
Shillingford et al. [44]	RNN	✓	3,886†	-	-	55.1
Ma et al. [26]	Transformer	✗	433	1,759	-	49.6*
Ma et al. [27]	Conformer	✓	438	-	-	46.9
AV-HuBERT-BASE [43]	Transformer	✗	30	1,759	-	46.1
SynthVSR	Conformer-BASE	✗	30	-	-	104.0
		✗	-	-	3,652	100.3
		✗	30	-	3,652	44.7
		✓	30	-	3,652	**43.3**

Table 1. Experimental results of low-resource labeled data setting on LRS3 (test). LM denotes whether or not a language model is used in the decoding. †Includes non-publicly available data. ‡Includes datasets that are only permitted for the purpose of academic research. hrs is an abbreviation for hours. *Result taken from [43].

- Using only 30 hours of LRS3 labelled data achieves **WER 43.3%**, outperforming the former methods using hundreds or thousands hours of data
- We show the first successful attempt that achieves an acceptable VSR WER with **only 30 hours of real data**

Experimental Results – LRS3 Benchmark Setting

Method	Backbone	LM	Labeled data (hrs)	Unlabeled data (hrs)	Synthetic data (hrs)	WER (%)
AV-HuBERT-BASE [43]	Transformer	✗	433	1,759	-	34.8
Makino et al. [30]	Transformer	✗	31,000†	-	-	33.6
Ma et al. [28]	Conformer	✓	1,459‡	-	-	31.5
Prajwal et al. [35]	Transformer	✓	2,676†	-	-	30.7
AV-HuBERT-LARGE [43]	Transformer	✗	433	1,759	-	28.6
AV-HuBERT-LARGE w. Self-Training [43]	Transformer	✗	433	1,759	-	26.9
Auto-AVSR [25]	Conformer	✓	3,448‡	-	-	19.1
Serdyuk et al. [42]	Transformer	✗	90,000†	-	-	25.9
Serdyuk et al. [41]	Transformer	✗	90,000†	-	-	17.0
SynthVSR	Conformer-BASE	✗	438	-	-	36.7
		✗	438	-	3,652	28.4
		✓	438	-	3,652	**27.9**
		✗	3,068	-	-	21.2
		✗	3,068	-	3,652	19.4
		✓	3,068	-	3,652	**18.7**
SynthVSR	Conformer-LARGE	✗	3,068	-	3,652	18.2
		✓	3,068	-	3,652	**16.9**

Table 2. Experimental results of LRS3 & high-resource labeled data setting on LRS3 (test). LM denotes whether or not a language model is used in the decoding. †Includes non-publicly available data. ‡Includes datasets that are only permitted for the purpose of academic research. hrs is an abbreviation for hours.

- Using only 438 hours of LRS3 labelled data achieves WER 27.9%, on-par with SOTA self-supervised method AV-HuBERT that uses external 1759 hours of unlabelled audio-visual data, but with fewer model parameters (250M vs 390M)

Experimental Results – High-Resource Setting

Method	Backbone	LM	Labeled data (hrs)	Unlabeled data (hrs)	Synthetic data (hrs)	WER (%)
AV-HuBERT-BASE [43]	Transformer	✗	433	1,759	-	34.8
Makino et al. [30]	Transformer	✗	31,000†	-	-	33.6
Ma et al. [28]	Conformer	✓	1,459‡	-	-	31.5
Prajwal et al. [35]	Transformer	✓	2,676†	-	-	30.7
AV-HuBERT-LARGE [43]	Transformer	✗	433	1,759	-	28.6
AV-HuBERT-LARGE w. Self-Training [43]	Transformer	✗	433	1,759	-	26.9
Auto-AVSR [25]	Conformer	✓	3,448‡	-	-	19.1
Serdyuk et al. [42]	Transformer	✗	90,000†	-	-	25.9
Serdyuk et al. [41]	Transformer	✗	90,000†	-	-	17.0
SynthVSR	Conformer-BASE	✗	438	-	-	36.7
		✗	438	-	3,652	28.4
		✓	438	-	3,652	**27.9**
		✗	3,068	-	-	21.2
		✗	3,068	-	3,652	19.4
		✓	3,068	-	3,652	**18.7**
SynthVSR	Conformer-LARGE	✗	3,068	-	3,652	18.2
		✓	3,068	-	3,652	**16.9**

Table 2. Experimental results of LRS3 & high-resource labeled data setting on LRS3 (test). LM denotes whether or not a language model is used in the decoding. †Includes non-publicly available data. ‡Includes datasets that are only permitted for the purpose of academic research. hrs is an abbreviation for hours.

- Using additional 2630 hours of ASR pseudo-labelled public audio-visual data, SOTA WER 16.9% is achieved on LRS3 with publicly-available data only, slightly surpassing the former SOTA method using 90k hours (29x more) of non-public labelled data.

Impact and Future Work

- We provide a scalable approach for VSR, reducing the need for large-scale annotated audio-visual data
- Potentially useful in low-resource language or new VSR applications (e.g., healthcare) where labeled data is scare
- Foster future work:
  - First benchmark with publicly-available synthetic data
  - How to better generate and leverage synthetic visual data?