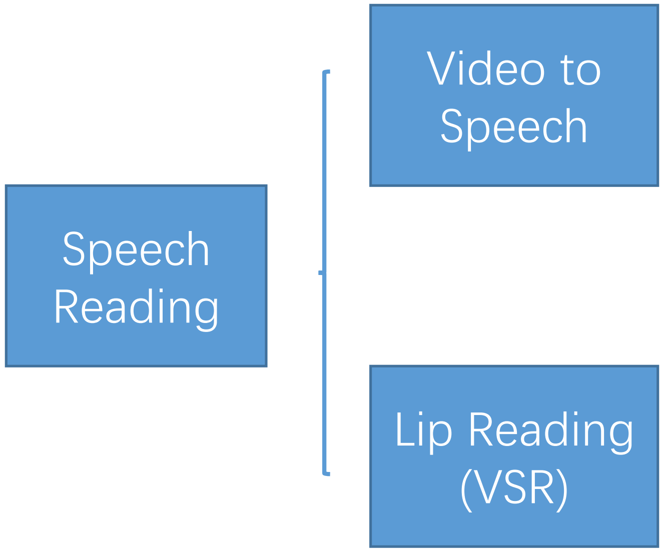
Recent Advance in VTS

Chen Chen

2022/08/05

Task: VTS

- Definition
  - Speechreading infers phonetic information from facial movements using visually observations.
  - Video-to-speech is the process of reconstructing the audio speech from a video of a spoken utterance.



Task: VTS

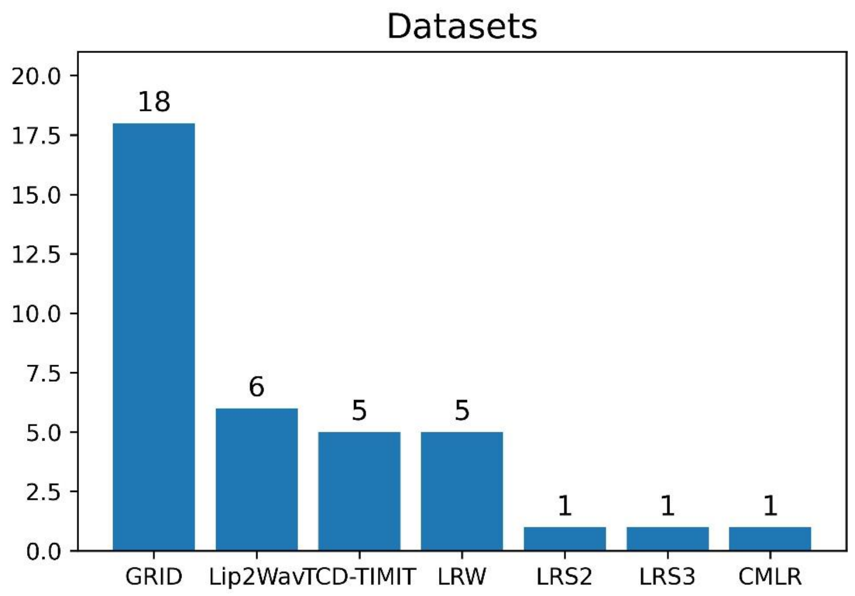
- Motivation
  - Can be done in a self-supervised manner
    - No text annotation required
  - Retain more identity information to enhance realism
    - VSR != VSR+TTS
- Target
  - Content
  - Indentity
- Difficulties
  - Weak information
  - Mismatch information
  - Noise information

Task: VTS

- Trend
  - Multi-stage
    - Visual feature → mel-spectrograms
    - mel-spectrograms → raw waveform
  - End-to-end
    - Raw video → Raw waveform
- Specific tasks
  - Single speaker vs. Multi-speaker
  - Seen speaker vs. Unseen speaker
  - Constrained vs. Unconstrained
    - view, resolution, light condition, vocabulary, ...

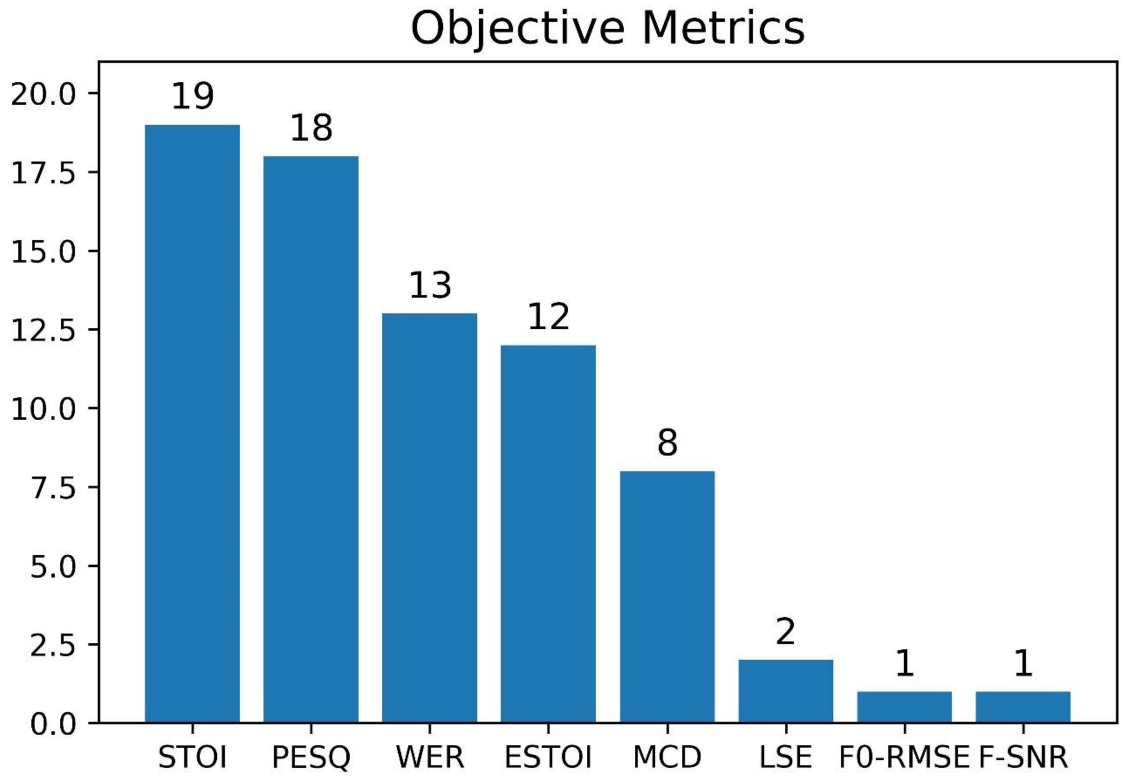
Datasets

- GRID
  - frontal view
  - small / close vocabulary
- TCD-TIMIT
  - frontal view
  - bigger vocabulary
- LRW
  - -30~30 view
  - 500 word class
- Lip2Wav
  - -90~90 view
  - large vocabulary
- LRS2
  - -30~30 view
  - large vocabulary
- LRS3
  - -90~90 view
  - large vocabulary
- CMLR
  - frontal view
  - large vocabulary
- LRW-1000
  - -90~90 view
  - 1000 word class



Metrics

- Intelligibility
  - Short-Time Objective Intelligibility (STOI)
  - Extended Short-Time Objective Intelligibility (ESTOI)
- Quality
  - Perceptual Evaluation of Speech Quality (PESQ)
  - Mean mel-cepstral distortion (MCD)
- Synchronisation (by SyncNet)
  - LSE-Confidence
  - LSE-Distance



Catalog

1. Speaker disentanglement in video-to-speech conversion
* 29th European Signal Processing Conference (EUSIPCO) 2021
* University POLITEHNICA of Bucharest | Technical University of Cluj-Napoca

2. LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading
* arxiv 2021
* University of Hamburg

3. Speech Reconstruction With Reminiscent Sound Via Visual Voice Memory
* IEEE/ACM Transactions on Audio, Speech, and Language Processing 2021
* Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

4. FastLTS: Non-Autoregressive End-to-End Unconstrained Lip-to-Speech Synthesis
* ACM MM 2022
* Zhejiang University, Hangzhou, China

5. Show Me Your Face, And I'll Tell You How You Speak
* arxiv 2022
* Saarland University

6. SVTS: Scalable Video-to-Speech Synthesis
* arxiv 2022
* Imperial College London, UK | University of Augsburg, Germany

Speaker disentanglement in video-to-speech conversion

* 29th European Signal Processing Conference (EUSIPCO) 2021

* University POLITEHNICA of Bucharest | Technical University of Cluj-Napoca

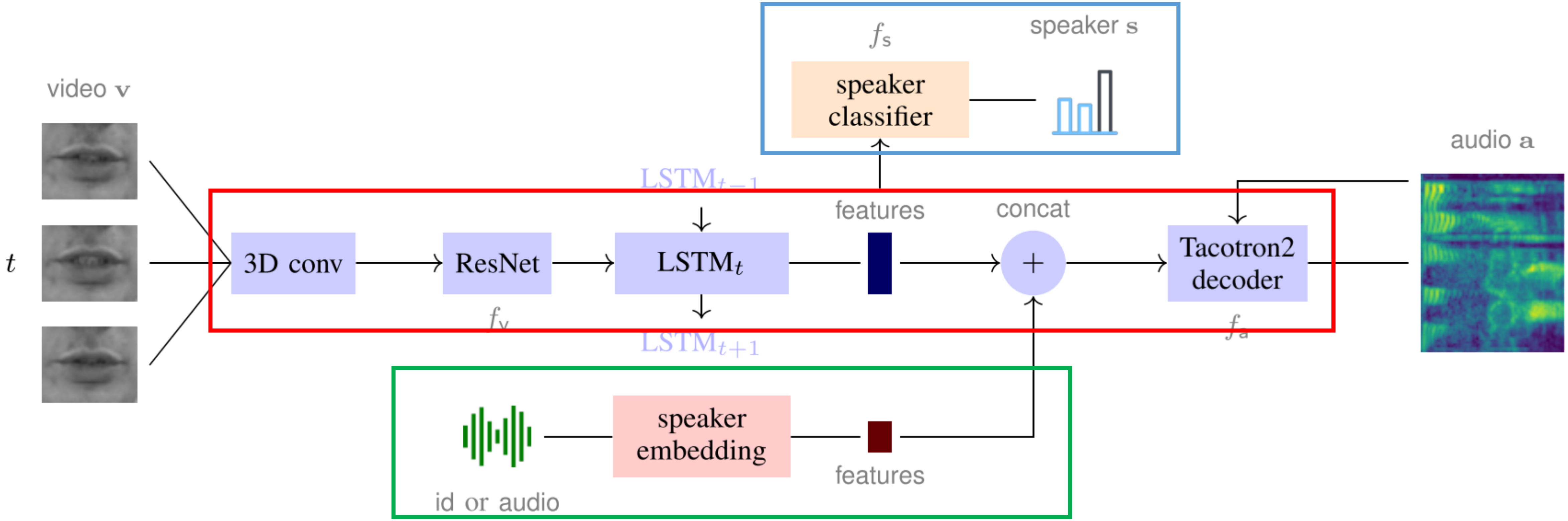
- Motivation
  - Leverage datasets with multiple speakers or few samples per speaker
  - Control speaker identity at inference time
- Dataset
  - GRID
- Tag
  - Constrained
  - Multi-speaker
  - Unseen speaker

Speaker disentanglement in video-to-speech conversion

* 29th European Signal Processing Conference (EUSIPCO) 2021

* University POLITEHNICA of Bucharest | Technical University of Cluj-Napoca

- Method
  - leverage state-of-theart systems from lip reading and text-to-speech synthesis
    - deep lip reading front-end as visual encoder
    - Tacotron2 architecture as speech decoder



- Method
  - Disentangling identity from content
    - Adversarial learning approach
      - discriminator learns to classify speakers based on visual features
      - generator changes the visual features to fool the discriminator and still be able to reconstruct the original audio

$$L_{\mathrm{d}}(f_{\mathsf{s}}) = H(\mathbf{s}, (f_{\mathsf{s}} \circ f_{\mathsf{v}})(\mathbf{v})))$$

$$L_{\mathrm{g}}(f_{\mathsf{a}}, f_{\mathsf{v}}) = \|\mathbf{a} - (f_{\mathsf{a}} \circ f_{\mathsf{v}})(\mathbf{v})\|_2^2 - \lambda H\left((f_{\mathsf{s}} \circ f_{\mathsf{v}})(\mathbf{v})\right)$$

    - Gradient reversal

$$\begin{aligned} L(f_{\mathsf{a}}, f_{\mathsf{s}}, f_{\mathsf{v}}) = \|\mathbf{a} - (f_{\mathsf{a}} \circ f_{\mathsf{v}})(\mathbf{v})\|_2^2 \\ + \lambda H(\mathbf{s}, (f_{\mathsf{s}} \circ f_{\mathsf{v}})(\mathbf{v}))) \end{aligned}$$

$\mathbf{v}$ denotes the input video
$\mathbf{a}$ the target audio
$\mathbf{s}$ the speaker identity
f_{v} the video processing net
f_{a} the audio decoder net
f_{s} the speaker classifier
H denotes the cross-entropy or entropy

Speaker disentanglement in video-to-speech conversion

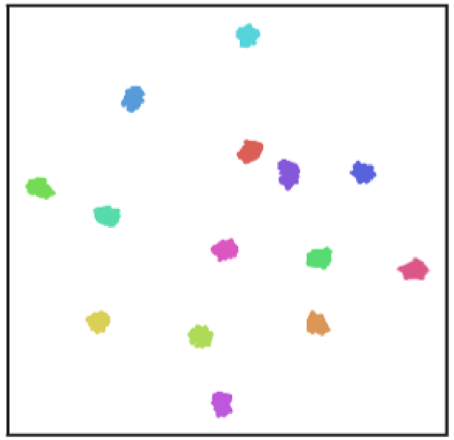
* 29th European Signal Processing Conference (EUSIPCO) 2021

* University POLITEHNICA of Bucharest | Technical University of Cluj-Napoca

- Experiment
  - B: speaker-independent baseline trained on all four speakers at once
  - B-spk: speaker-dependent baseline trained for each speaker separately
  - SI: model trained on all four speakers at once with speaker identity

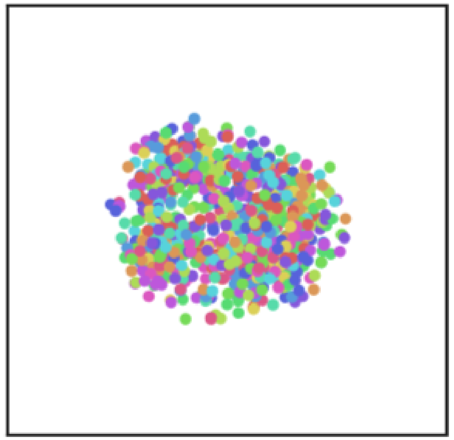
	STOI ↑	PESQ ↑	MCD ↓	WER ↓
Lip2AudSpec [3]	0.446	1.82	38.14	32.5
V2S GAN [5]	0.518	1.71	22.29	26.6
V2S GAN [5]†	**0.525**	1.72	**22.02**	27.1
B	0.470	**1.88**	32.28	21.8
B-spk	0.452	1.82	32.42	**17.8**
SI	0.468	1.85	32.08	19.9

Speaker embeddings for synthesised audio generated



speaker independent

B



speaker dependent

fixed speaker embedding

SI

Speaker disentanglement in video-to-speech conversion

* 29th European Signal Processing Conference (EUSIPCO) 2021

* University POLITEHNICA of Bucharest | Technical University of Cluj-Napoca

• Experiment

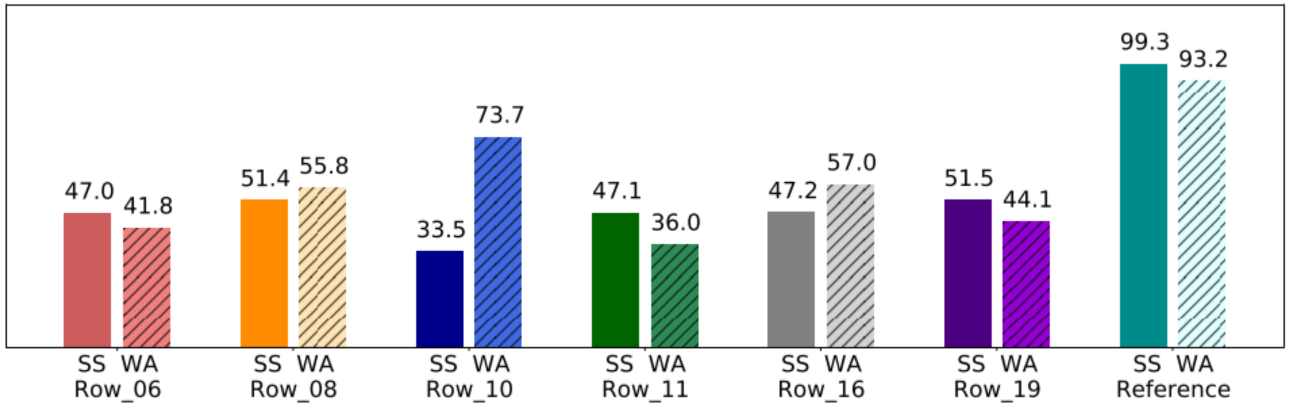
• Objective

	Architecture	Drop	Disentanglement		WER ↓	EER ↓
1	V2S GAN [5]	–	–	–	41.9	N/A
2	B	no	–	–	41.9	N/A
3	SI	no	–	–	43.7	6.9
4		yes	–	–	43.8	7.1
5		yes	dispel	MLP	50.2	7.5
6		yes	dispel	linear	43.7	**6.8**
7		yes	rev. grad.	MLP	45.2	6.9
8		yes	rev. grad.	linear	**42.7**	7.3
9	SE	no	–	–	36.5	18.0
10		yes	–	–	**31.2**	48.6
11		yes	dispel	MLP	41.9	**7.1**
12		yes	dispel	linear	35.5	12.7
13		yes	rev. grad.	MLP	37.7	8.9
14		yes	rev. grad.	linear	36.1	13.6
15	SE-norm	no	–	–	40.6	11.7
16		yes	–	–	**38.7**	12.5
17		yes	dispel	MLP	49.6	7.8
18		yes	dispel	linear	40.1	10.6
19		yes	rev. grad.	MLP	41.5	**7.6**
20		yes	rev. grad.	linear	38.9	11.9

• Subjective

SS: speaker similarity

WA: intelligibility, evaluated in terms of word accuracy



demo: https://speed.pub.ro/xts/

LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading

* arxiv 2021

* University of Hamburg

• Motivation

- investigate the impact of crossmodal self-supervised pre-training for speech reconstruction (video-to-audio) by leveraging the natural co-occurrence of audio and visual streams in videos

• Dataset

- GRID
- TCD-TIMIT
- CMLR

Language	Dataset	#Spk.	#Utt.	#Vocab.	#hours	Usage	Modality
Multi-Language	VoxCeleb2 [77]	6112	1.1M	-	2442	LipSound2 pre-training	Audio-Visual
English	GRID [78]	51	33k	51	27.5	LipSound2 fine-tuning	
	TCD-TIMIT [32]	59	5.4k	5.9k	7		
	LJSpeech [74]	1	13.1k	-	24	WaveGlow training	Audio
	LibriSpeech [79]	2484	292.3k	-	960	Acoustic model pre-training	
Chinese	CMLR [80]	11	102k	3.5k	87.7	LipSound2 fine-tuning	Audio-Visual
	AISHELL-2 [81]	1991	-	-	1000	Acoustic model pre-training	Audio

• Tag

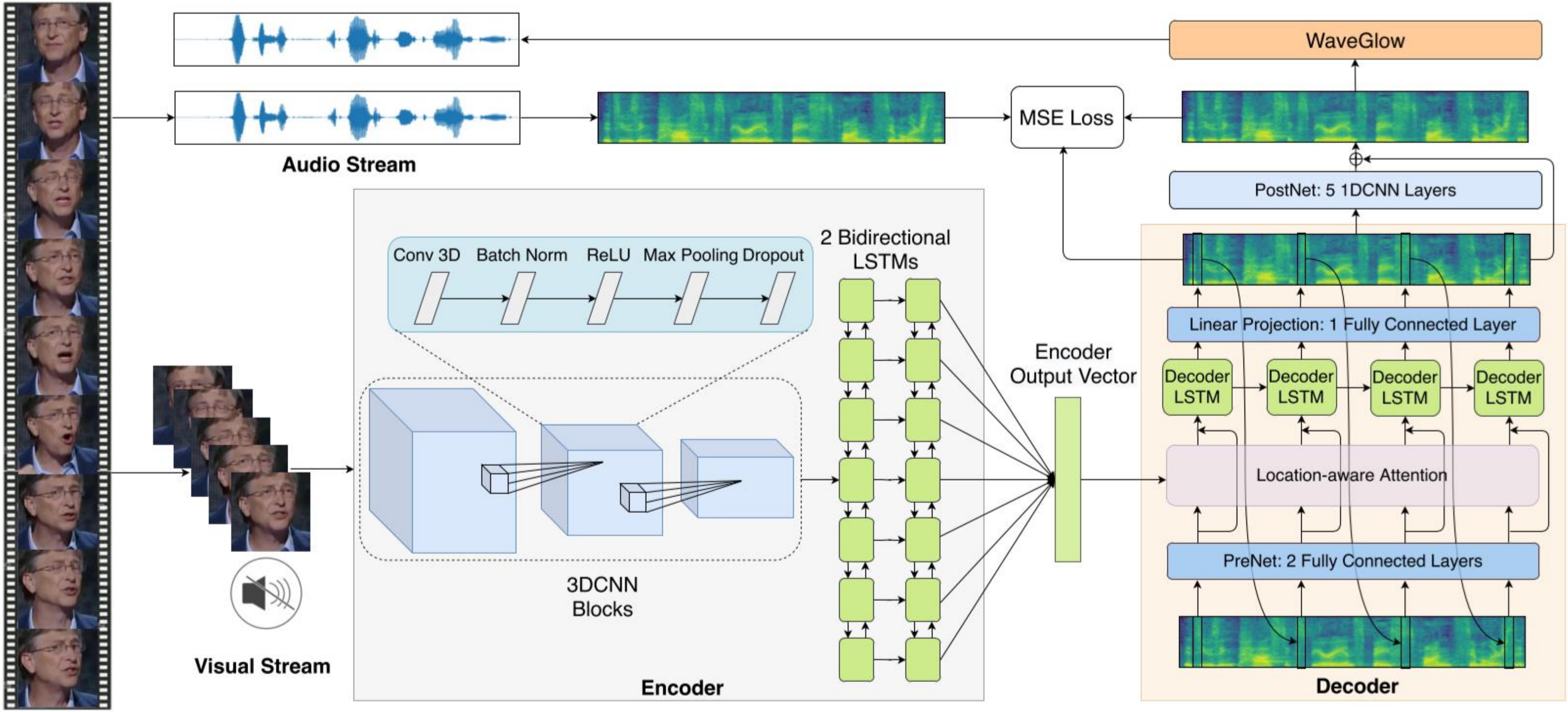
- Unconstrained
- Multi-speaker
- Unseen speaker

LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading

* arxiv 2021

* University of Hamburg

• Method



LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading

* arxiv 2021

* University of Hamburg

- Method
  - Pre-training
    - on VoxCeleb2
  - Fine-tuning
    - on GRID / TCD-TIMIT / CMLR

LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading

* arxiv 2021
* University of Hamburg

- Experiment on VTS
  - Speaker-dependent
    - Multi-speaker on GRID (Speaker S1 −S4) / TCD-TIMIT (Lipspeaker 1 − 3)

	GRID		TCD-TIMIT	
Model	ESTOI	PESQ	ESTOI	PESQ
Vid2Speech [18]	0.335	1.734	0.298	1.136
Lip2AudSpec [19]	0.352	1.673	0.316	1.254
Vougioukas et al. [34]	0.361	1.684	0.321	1.218
Ephrat et al. [31]	0.376	1.825	0.310	1.231
Lip2Wav [36]	0.535	1.772	0.365	1.350
vid2voc-M-VSR [37]	0.455	1.900	-	-
LipSound2	**0.592**	**2.328**	**0.372**	**1.490**

  - Speaker-independent
    - Multi-speaker on GRID / TCD-TIMIT

	GRID		TCD-TIMIT	
Model	ESTOI	PESQ	ESTOI	PESQ
Vougioukas et al. [34]	0.198	1.24	-	-
vid2voc-M-VSR [37]	0.227	1.23	-	-
vid2voc-F-VSR [37]	0.210	1.25	-	-
LipSound2	**0.363**	**1.72**	**0.30**	**1.31**

LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading

* arxiv 2021

* University of Hamburg

- Experiment on VTS
  - Speaker-dependent
    - Multi-speaker on CMLR
  - Speaker-independent
    - Test on Speaker S1, S6, Train on others

	Speaker-dependent		Speaker-independent	
Model	ESTOI	PESQ	ESTOI	PESQ
LipSound2	**0.36**	**1.43**	**0.28**	**1.21**

demo: https://leyuanqu.github.io/LipSound2/

LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading

* arxiv 2021

* University of Hamburg

- Experiment on VSR
    - Video → Audio → Text

	GRID		TCD-TIMIT	
Model	Spk-Dep	Spk-Indep	Spk-Dep	Spk-Indep
Audio Gold Standard	22.36	21.88	15.86	15.21
+Fine-tuning	0.15	0.35	5.42	6.73
LipNet [43]	5.6	13.6	-	-
LipNet+LM [43]	4.8	11.4	-	-
PCPG+LM [87]	-	11.2	-	-
TVSR-Net [88]	-	9.1	-	-
WAS [2]	3.0	-	-	-
LCANet[89]	2.9	-	-	-
DualLip [90]	2.7	-	-	-
LipSound [20]	2.5	-	-	-
CD-DNN [86]	-	-	51.26	57.03
MobiLipNetV2 [91]	-	-	-	53.01
LipSound2	1.9	7.3	41.37	46.29
LipSound2 + LM	**1.5**	**6.4**	**39.77**	**43.53**

Model	Spk-dep	Spk-indep
Audio Gold Standard	19.25	16.2
+Fine-tuning	3.88	4.89
WAS [2]	38.93	-
CSSMCM [80]	32.48	-
LIBS [92]	31.27	-
LipSound2	25.03	36.56
LipSound2 + LM	**22.93**	**33.44**

Speech Reconstruction With Reminiscent Sound Via Visual Voice Memory

* IEEE/ACM Transactions on Audio, Speech, and Language Processing 2021

* Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

- Motivation
  - Reconstruct speech from silent video, in both speaker dependent and independent ways
- Dataset
  - GRID
  - Lip2Wav
- Tag
  - Unconstrained
  - Unseen speaker
  - Multi-speaker

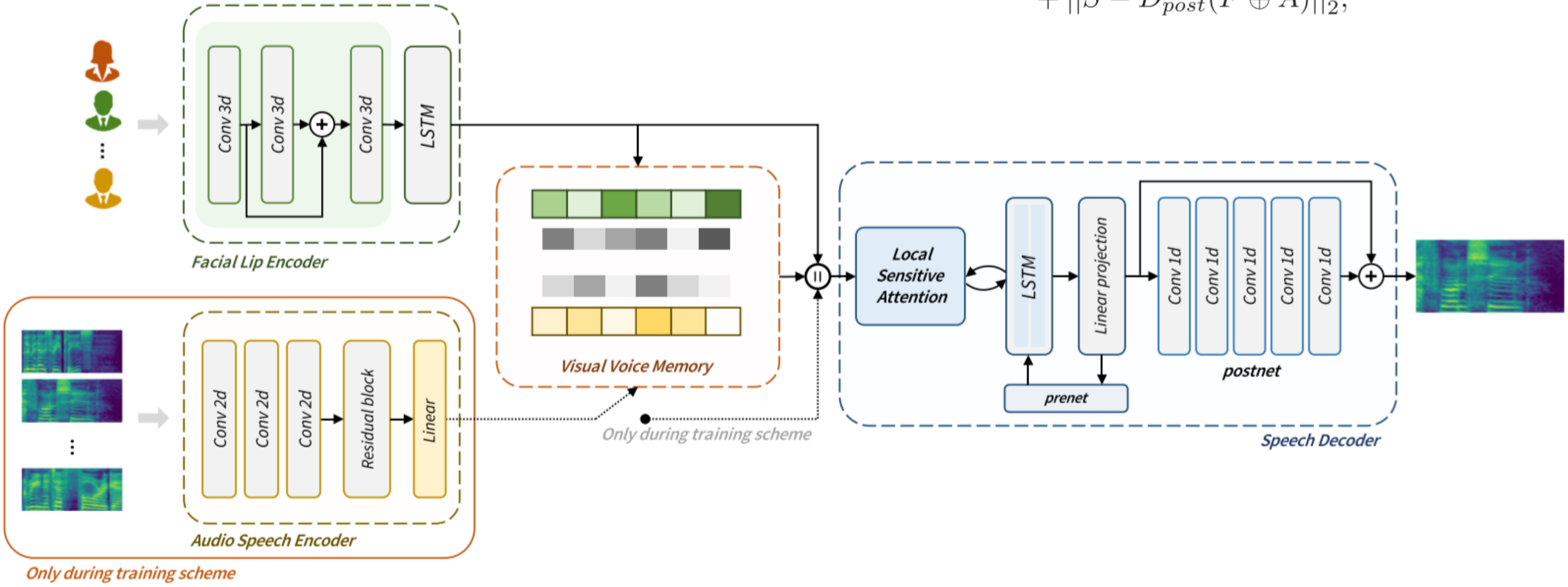
Speech Reconstruction With Reminiscent Sound Via Visual Voice Memory

* IEEE/ACM Transactions on Audio, Speech, and Language Processing 2021

* Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

• Method

$$\mathcal{L}_{mel,\bar{A}} = ||S - D_{pre}(F \oplus \bar{A})||_2^2 + ||S - D_{post}(F \oplus \bar{A})||_2^2,$$

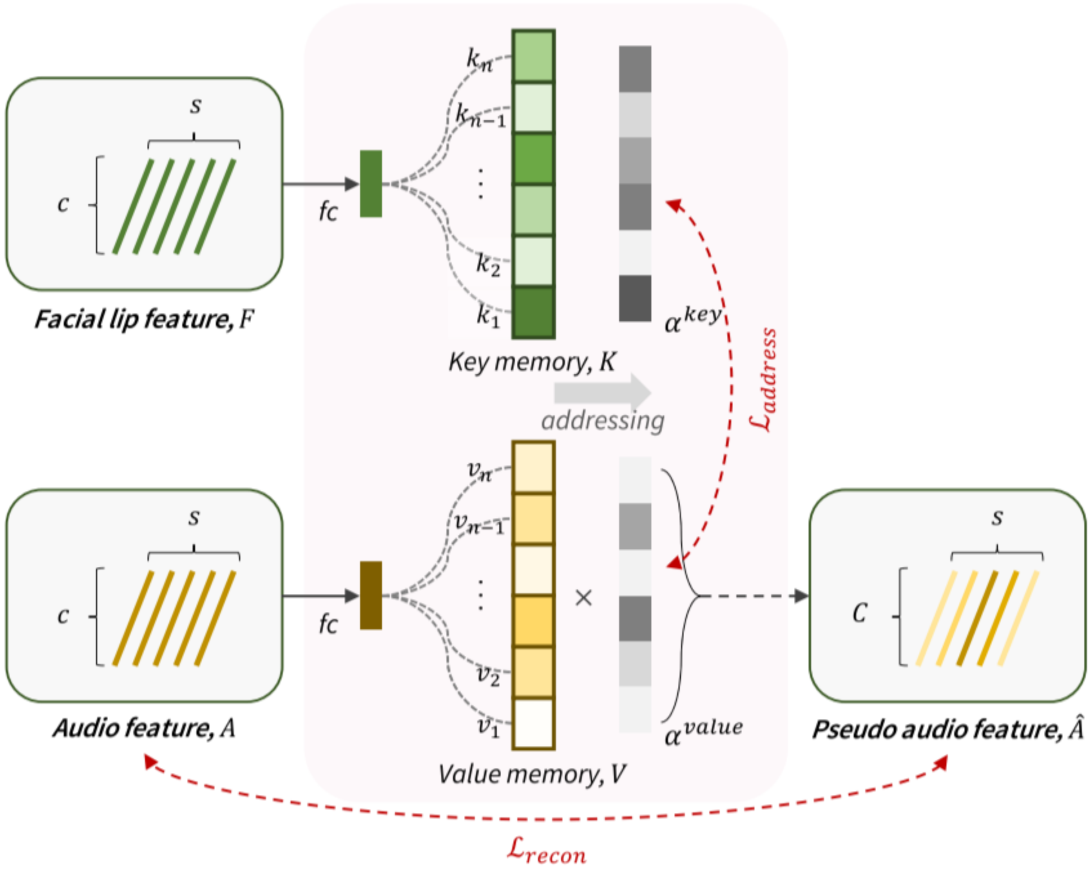


Speech Reconstruction With Reminiscent Sound Via Visual Voice Memory

* IEEE/ACM Transactions on Audio, Speech, and Language Processing 2021

* Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

- Method



(a) Training objective functions in Visual Voice Memory

$$\mathcal{L}_{recon} = ||A - \hat{A}||_2^2.$$

$$\mathcal{L}_{address} = D_{KL}(\alpha^{value}||\alpha^{key}).$$

$$\mathcal{L}_{mel,\bar{A}} = ||S - D_{pre}(F \oplus \bar{A})||_2^2 + ||S - D_{post}(F \oplus \bar{A})||_2^2,$$

$$\mathcal{L}_{mel} = L_{mel,\bar{A}} + L_{mel,A}.$$

Speech Reconstruction With Reminiscent Sound Via Visual Voice Memory

* IEEE/ACM Transactions on Audio, Speech, and Language Processing 2021
* Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

• Experiment

• Speaker Dependent

Method	Speaker	STOI	ESTOI	PESQ
Ephrat et al. [24]	*Chemistry Lectures*	0.165	0.087	1.056
GAN-based [32]		0.192	0.132	1.057
Lip2Wav [26]		0.416	0.284	1.300
Proposed model		**0.566**	**0.429**	**1.529**
Ephrat et al. [24]	*Chess Analysis*	0.184	0.098	1.139
GAN-based [32]		0.195	0.104	1.165
Lip2Wav [26]		0.418	0.292	1.400
Proposed model		**0.506**	**0.334**	**1.503**
Ephrat et al. [24]	*Deep Learning*	0.112	0.043	1.095
GAN-based [32]		0.144	0.070	0.121
Lip2Wav [26]		0.282	0.183	**1.671**
Proposed model		**0.576**	**0.402**	1.612
Ephrat et al. [24]	*Hardware Security*	0.192	0.064	1.043
GAN-based [32]		0.251	0.110	1.035
Lip2Wav [26]		0.446	0.311	1.290
Proposed model		**0.504**	**0.337**	**1.366**
Ephrat et al. [24]	*Ethical Hacking*	0.143	0.064	1.065
GAN-based [32]		0.171	0.089	1.079
Lip2Wav [26]		0.369	0.220	**1.367**
Proposed model		**0.463**	**0.304**	1.362

• Multi-speaker

Dataset	Method	STOI	ESTOI	PESQ	WER (%)
GRID	Lip2Wav [26]	0.707	0.530	1.715	21.33
	Proposed	**0.754**	**0.602**	**2.112**	**9.83**
Lip2Wav	Lip2Wav [26]	0.404	0.205	1.356	-
	Proposed	**0.496**	**0.281**	**1.537**	-

• Speaker Independent

Method	STOI	ESTOI	PESQ	WER(%)
GAN-based [32]	0.445	-	1.240	40.50
Lip2Wav [26]	0.565	0.279	1.279	38.37
Proposed model	**0.600**	**0.315**	**1.332**	**37.96**

https://github.com/joannahong/Speech-Reconstruction-with-Reminiscent-Sound-via-Visual-Voice-Memory

Speech Reconstruction With Reminiscent Sound Via Visual Voice Memory
* IEEE/ACM Transactions on Audio, Speech, and Language Processing 2021
* Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

- Experiment
  - Ablation study on memory slot size

Dataset	Memory slot size	STOI	ESTOI	PESQ	WER (%)
GRID	0 (Baseline)	0.707	0.530	1.715	21.33
	50	0.749	0.591	2.080	12.67
	150	0.749	0.595	2.076	11.33
	360	**0.754**	**0.602**	**2.112**	**9.83**
Lip2Wav	0 (Baseline)	0.404	0.205	1.356	-
	50	0.471	0.256	1.517	-
	150	0.487	0.267	1.510	-
	360	**0.496**	**0.281**	**1.537**	-

  - effectiveness on mel-spectrum resemble

MEAN COSINE SIMILARITY OF GRID TEST SET BETWEEN THE ORIGINAL AUDIO FEATURE A AND THE IMPRINTED AUDIO FEATURES $\bar{A}$. *MEM. ADDR.* REFERS TO MEMORY ADDRESSING VECTORS

Method	Cosine similarity
Proposed Model	**0.704**
- top 1 mem. addr	0.568
- top 5 mem. addr	-0.070
- top 10 mem. addr	-0.538

  - speech generation performances based on variety of the corrupted addressing vectors

Method	STOI	ESTOI	PESQ	WER (%)
Proposed model	**0.754**	**0.602**	**2.112**	**9.83**
- top 1 mem. addr.	0.673	0.508	1.786	19.42
- top 5 mem. addr.	0.630	0.446	1.557	34.25
- top 10 mem. addr.	0.633	0.447	1.579	36.58

FastLTS: Non-Autoregressive End-to-End Unconstrained Lip-to-Speech Synthesis

* ACM MM 2022

* Zhejiang University, Hangzhou, China

- Motivation
    - Two-stage pipeline causes cumbersome deployment and degradation of speech quality due to error propagation
    - Autoregressive model suffers from high inference latency, flow-based model has high memory occupancy

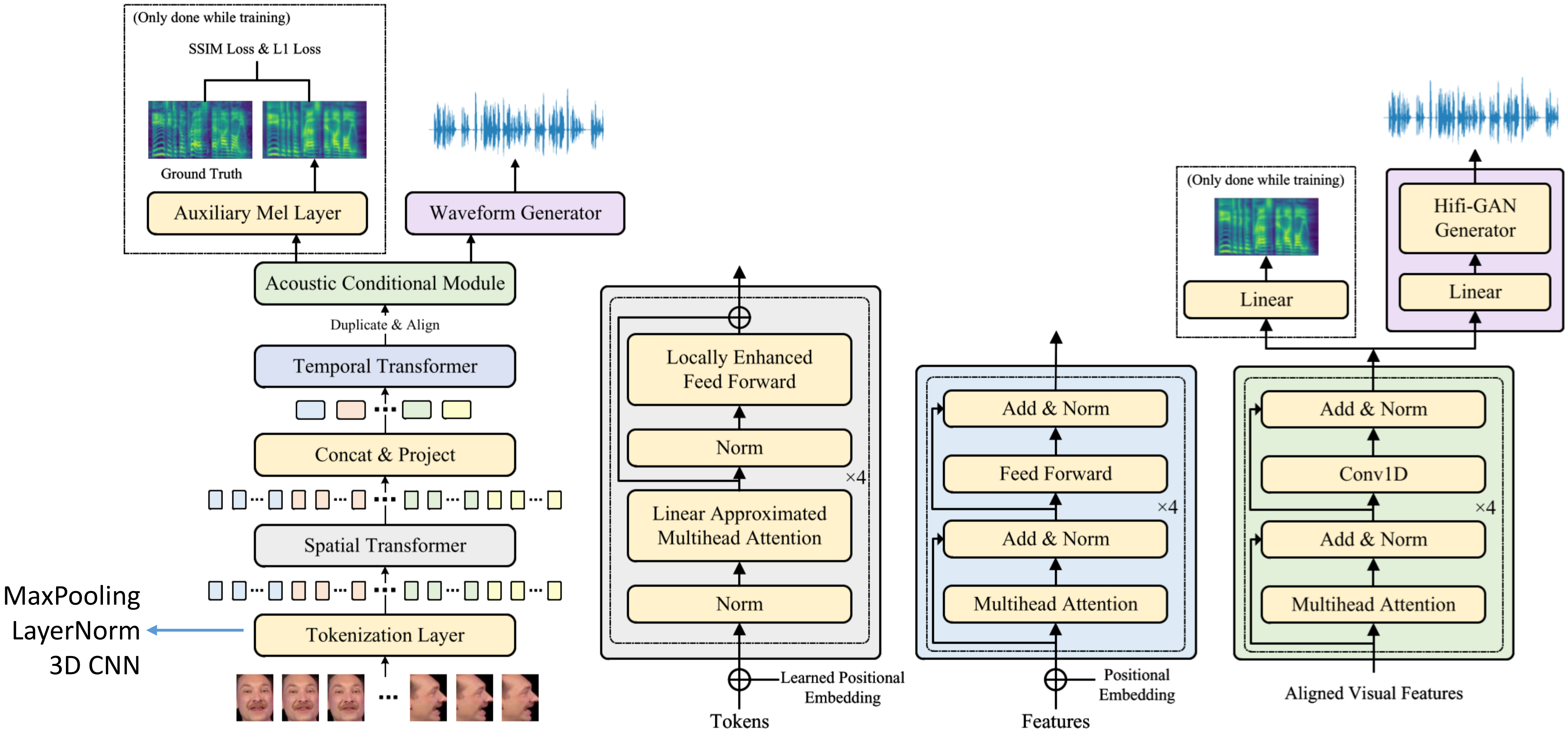
- Dataset
    - Lip2Wav
    - GRID
- Tag
    - Unconstrained
    - Seen speaker
    - Single speaker

FastLTS: Non-Autoregressive End-to-End Unconstrained Lip-to-Speech Synthesis

* ACM MM 2022

* Zhejiang University, Hangzhou, China

- Methods



(a) overall architecture (b) spatial transformer (c) temporal transformer (d) audio decoder

FastLTS: Non-Autoregressive End-to-End Unconstrained Lip-to-Speech Synthesis

* ACM MM 2022

* Zhejiang University, Hangzhou, China

- Methods
  - Visual Encoder
    - Tokenization layer
      - used to preliminarily extract local features and produce spatio-temporal tokens for the transformer
    - Spatial transformer
      - used to model the correlation among spatially adjacent tokens, and only calculates attentions on tokens extracted from the same temporal index
      - use linear approximation of self-attention proposed in Performer to reduces the computation burden of self-attention
      - employ the Locally Enhanced Feed Forward network which does convolution on the depth dimension of the features
    - Temporal transformer
      - models the temporal correlation between the hiddens
  - Acoustic Conditional Module
    - turns visual features into acoustic features
    - simply duplicate the visual features for alignment

FastLTS: Non-Autoregressive End-to-End Unconstrained Lip-to-Speech Synthesis

* ACM MM 2022
* Zhejiang University, Hangzhou, China

• Methods

• Two stage Training Method

- train the visual encoder and the acoustic conditional module

$$\mathcal{L}_{\text{SSIM}} = \frac{1}{L_{mel}} \sum_{n=1}^{L_{mel}} 1 - \text{SSIM}(y_n, \hat{y}_n) \qquad \mathcal{L}_{L1} = \frac{1}{L_{mel}} \sum_{n=1}^{L_{mel}} \|y_n - \hat{y}_n\|_1$$

$$\mathcal{L}_{stage1} = \lambda_{\text{SSIM}} \mathcal{L}_{\text{SSIM}} + \lambda_{L1} \mathcal{L}_{L1}$$

- plug the waveform generator into the model, freeze visual encoder and acoustic conditional module

$$\mathcal{L}_{adv}(D;G) = \mathbb{E}_{(x,s)}\left[(D(x)-1)^2 + (D(G(s)))^2\right] \qquad \mathcal{L}_{adv}(G;D) = \mathbb{E}_{(x,s)}\left[(D(G(s))-1)^2\right]$$

$$\mathcal{L}_{mel}(G) = \mathbb{E}_{(x,s)}\left[\|\phi(x) - \phi(G(s))\|_1\right] \qquad \mathcal{L}_{FM}(G;D) = \mathbb{E}_{(x,s)}\left[\sum_{i=1}^{T} \frac{1}{N_i} \left\|D^i(x) - D^i(G(s))\right\|_1\right]$$

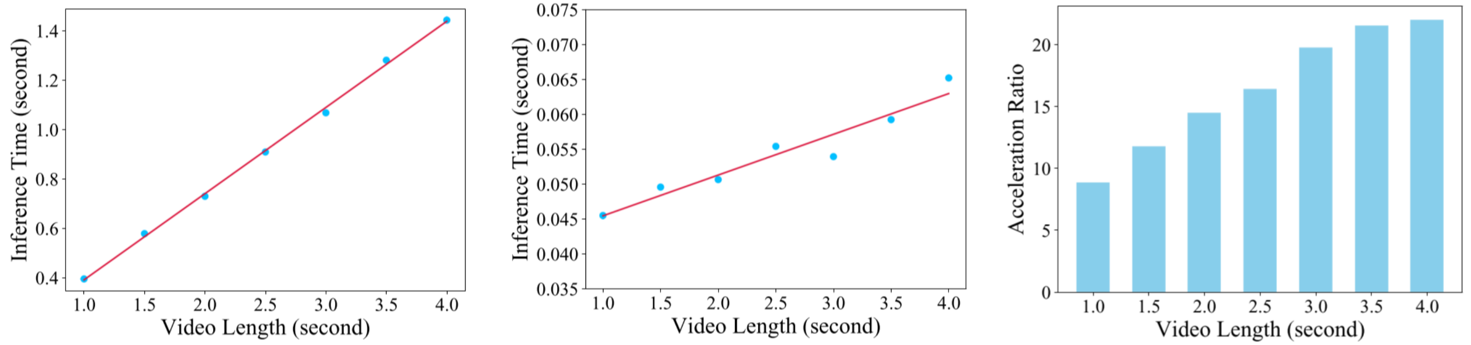
$$\mathcal{L}_{stage2-G} = \lambda_a \mathcal{L}_{adv}(G;D) + \lambda_m \mathcal{L}_{mel}(G) + \lambda_f \mathcal{L}_{FM}(G;D)$$

$$\mathcal{L}_{stage2-D} = \mathcal{L}_{adv}(D;G)$$

FastLTS: Non-Autoregressive End-to-End Unconstrained Lip-to-Speech Synthesis

* ACM MM 2022
* Zhejiang University, Hangzhou, China

- Experiments



(a) Wav Inference Time of Lip2Wav (b) Wav Inference Time of FastLTS (c) Acceleration Ratio over Time

Table 1: MOS on Lip2Wav Dataset

Speaker	Method	Quality	Intelli.	Natural.
Chess Analysis	Lip2Wav	3.53 ± 0.10	3.51 ± 0.09	3.48 ± 0.09
	FastLTS	3.79 ± 0.09	3.82 ± 0.10	3.59 ± 0.08
	GT	4.08 ± 0.07	3.91 ± 0.08	4.10 ± 0.06
Chemistry Lectures	Lip2Wav	3.60 ± 0.10	3.88 ± 0.10	3.78 ± 0.09
	FastLTS	3.84 ± 0.10	3.73 ± 0.11	3.87 ± 0.09
	GT	4.06 ± 0.08	3.90 ± 0.09	4.10 ± 0.07
Hardware Security	Lip2Wav	3.67 ± 0.10	3.57 ± 0.11	3.74 ± 0.11
	FastLTS	3.86 ± 0.12	3.89 ± 0.13	3.80 ± 0.14
	GT	3.94 ± 0.10	4.01 ± 0.10	4.02 ± 0.09

Table 3: MOS on GRID Dataset

Method	Quality	Intelligibility	Naturalness
Lip2Wav	3.27 ± 0.11	3.47 ± 0.13	3.54 ± 0.12
FastLTS	3.59 ± 0.09	3.68 ± 0.09	3.73 ± 0.08
GT	3.60 ± 0.10	3.76 ± 0.09	3.78 ± 0.10

Table 4: PESQ on GRID dataset

Method	PESQ
Vid2Speech [10]	1.734
Lip2AudSpec [1]	1.673
GAN-based [37]	1.684
Ephrat et al. [9]	1.825
Lip2Wav [24]	1.772
VAE-based [41]	1.932
Vocoder-based [21]	1.900
VCA-GAN [18]	**2.008**
FastLTS	1.939

Table 5: Parameter Amounts of Three Different Models

Model	Parameters	Relative Size
Autoregressive Model		
Lip2Wav	39.87M	1.00×
Non-autoregressive Models		
GlowLTS	85.92M	2.16×
FastLTS(ours)	50.09M	1.26×

Table 6: CMOS Comparison in Ablation Studies

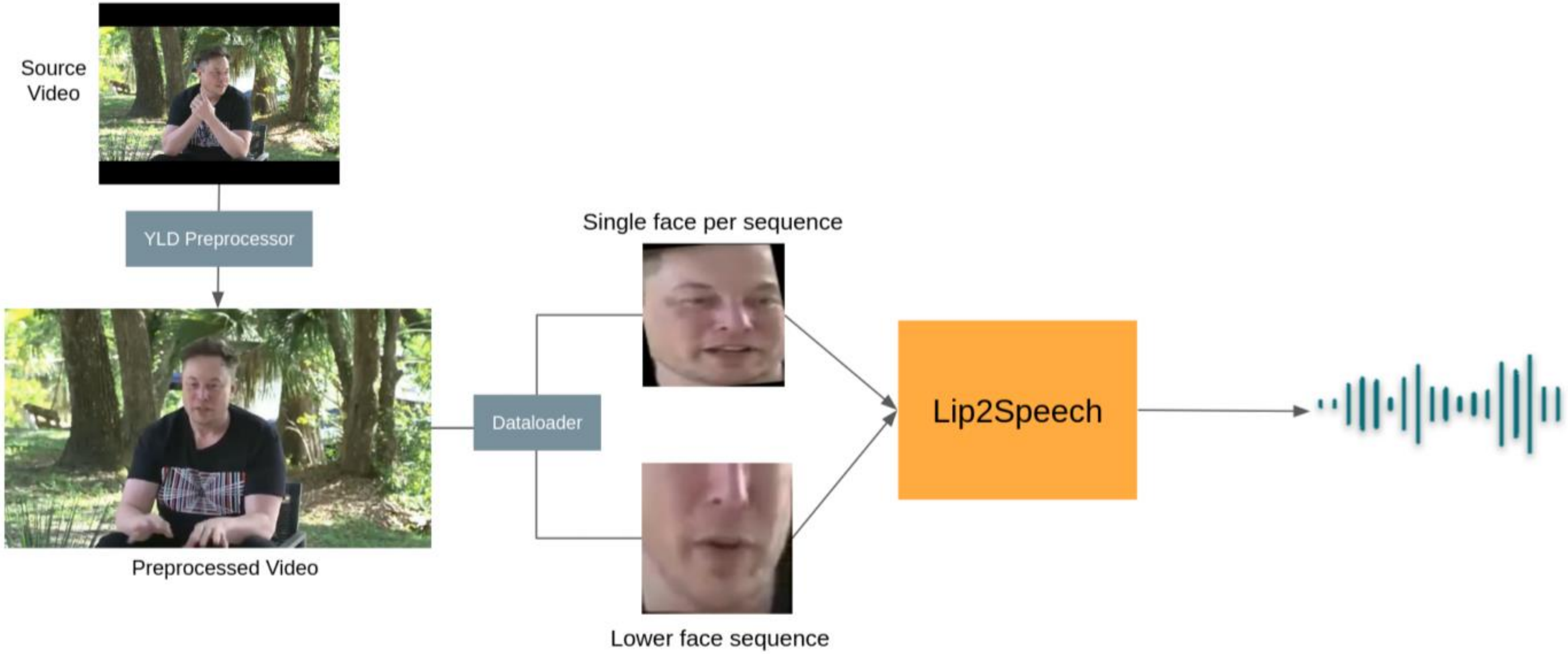
Method	Quality	Intelli.	Natural.
FastLTS	0	0	0
w/o waveform generator	-0.274	-0.075	-0.103
w/o conditional module	N/A	N/A	N/A
w/o first training stage	N/A	N/A	N/A

Show Me Your Face, And I'll Tell You How You Speak

* arxiv 2022

* Saarland University

- Motivation
  - Capture the speaker's voice identity through their facial characteristics and condition them along with the lip movements to generate speaker identity aware speech
- Dataset
  - LRW
  - AVSpeech
  - UTKFace
  - YLD
- Tag
  - Unconstrained
  - Unseen Speaker
  - Multi-speaker

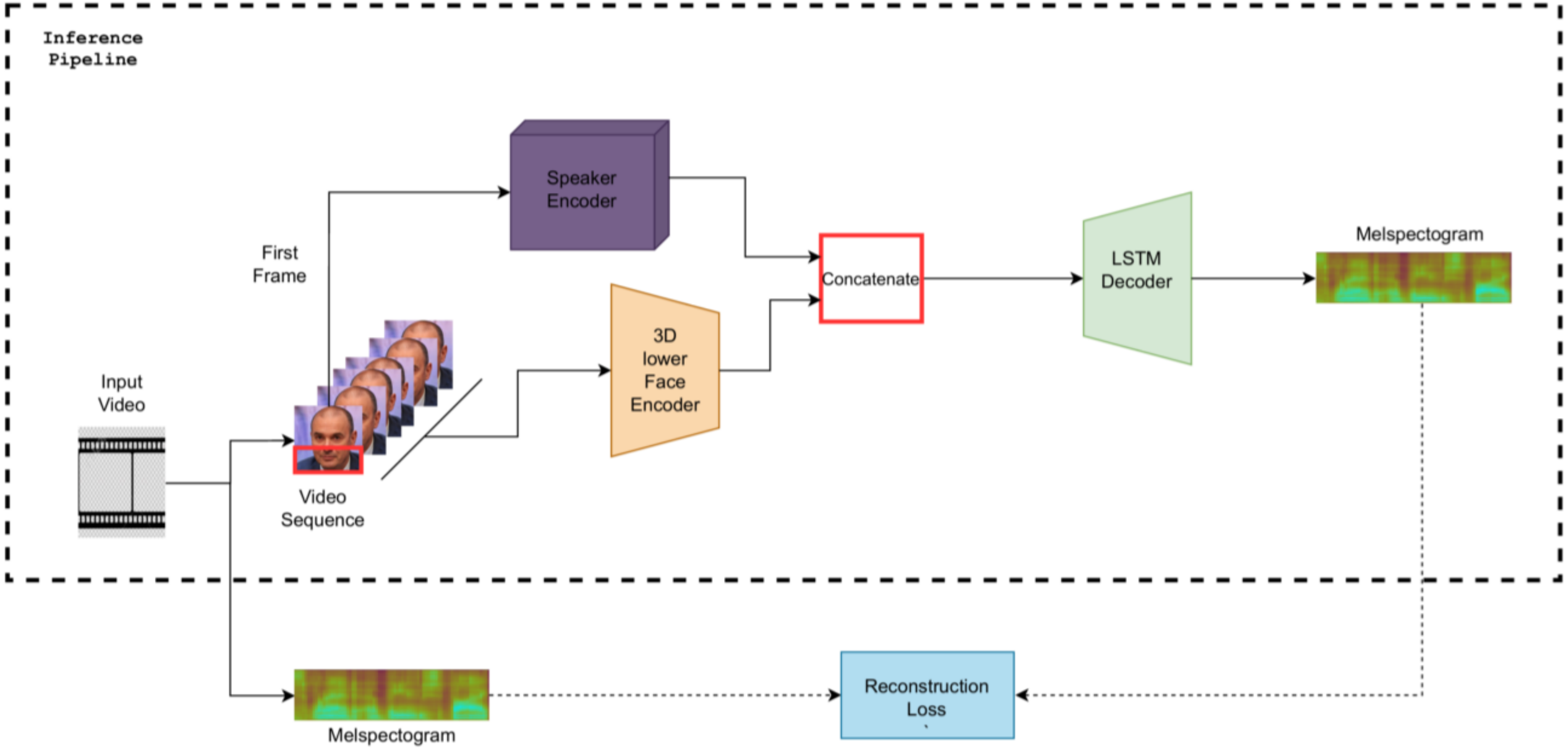


Show Me Your Face, And I’ll Tell You How You Speak

* arxiv 2022

* Saarland University

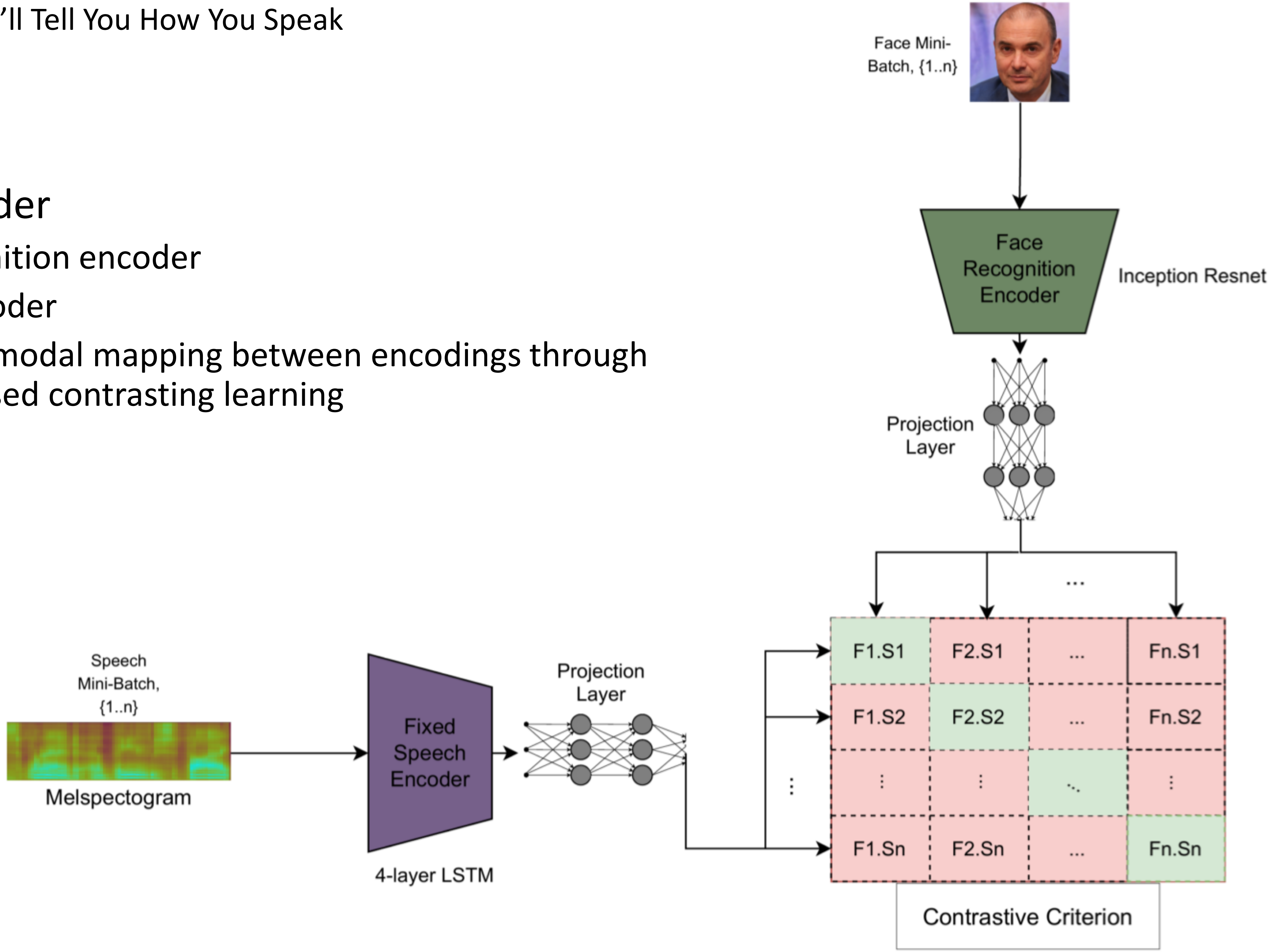
- Methods
  - Speaker Encoder
  - Face Encoder
  - LSTM Decoder
  - MSE loss on melspectogram



Show Me Your Face, And I'll Tell You How You Speak

* arxiv 2022
* Saarland University

- Methods
  - Speaker Encoder
    - Face Recognition encoder
    - Speech encoder
    - learn cross-modal mapping between encodings through instance based contrasting learning



Show Me Your Face, And I’ll Tell You How You Speak

* arxiv 2022

* Saarland University

- Methods
  - Face Encoder
    - 3D conv -> Shufflenet 2D conv

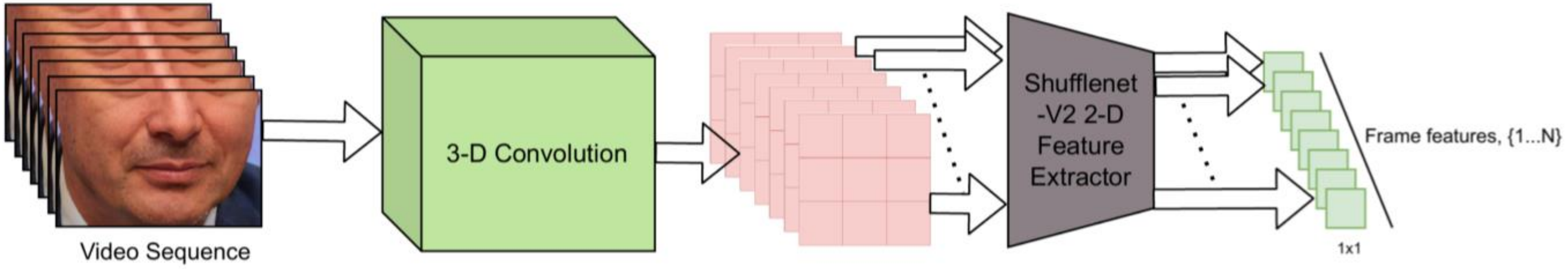


Fig. 4. The video feature extractor encodes the frame sequence by applying spatio-temporal convolutions.

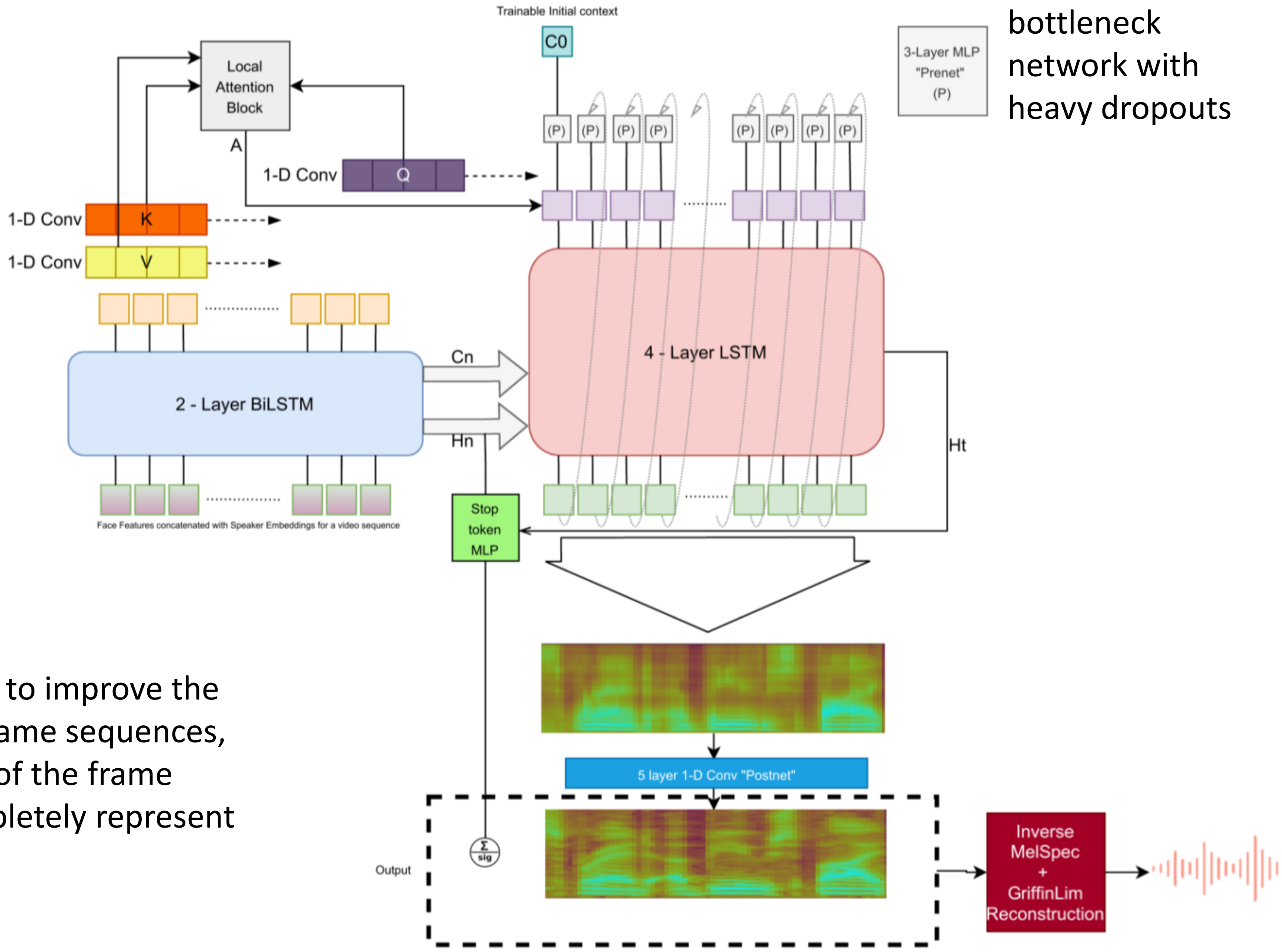
Show Me Your Face, And I'll Tell You How You Speak

* arxiv 2022

* Saarland University

- Methods
  - LSTM Decoder
    - BiLSTM encoder
    - Localized attention
    - LSTM decoder

bottleneck network with heavy dropouts



use localized attention mechanism to improve the contextual information from the frame sequences, as the condensed latent encoding of the frame sequences will not be able to completely represent the temporal semantic flow

Show Me Your Face, And I'll Tell You How You Speak

* arxiv 2022
* Saarland University

- Experiment
  - Speaker Encoder

Table 1. Voices generated with different embeddings and their respective MOS score compared to the ground truth.

Voice	Quality
ground truth	**4.56**
speaker audio embedding	3.37
speaker face embedding	3.55

Voice	Correlation
ground truth	**4.44**
speaker audio embedding	3.12
speaker face embedding	3.03

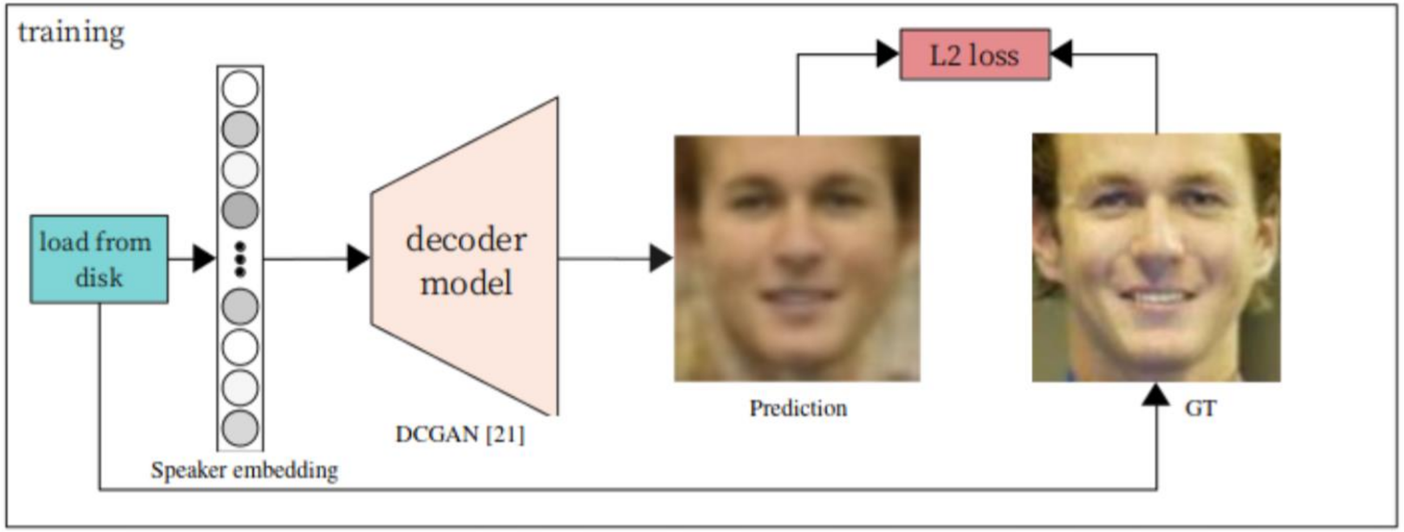


Fig. 7. Left is the ground truth and right is the reconstructed face.

Show Me Your Face, And I'll Tell You How You Speak

* arxiv 2022

* Saarland University

- Experiment
  - Lip2Speech

Table 2. Quantitative results of our model on the 153 test samples

	STOI ↑	ESTOI ↑	PESQ ↑	WER ↓
Lip2Speech	1.38	0.66	0.42	26.1%

Table 3. Quantitative results of other models on LRW test split

	STOI ↑	ESTOI ↑	PESQ ↑	WER ↓
Lip2Wav	0.543	0.344	1.197	34.3%
Chung et al.	NA	NA	NA	38.8%

SVTS: Scalable Video-to-Speech Synthesis

* arxiv 2022
* Imperial College London, UK | University of Augsburg, Germany

- Motivation
  - Leverage massive amount of audio-visual data
  - Propose training procedures which can easily scale to very large datasets
- Dataset
  - GRID
  - LRW
  - LRS3
  - VoxCeleb2
- Tag
  - Unconstrained
  - Unseen Speaker
  - Multi-speaker

SVTS: Scalable Video-to-Speech Synthesis

* arxiv 2022

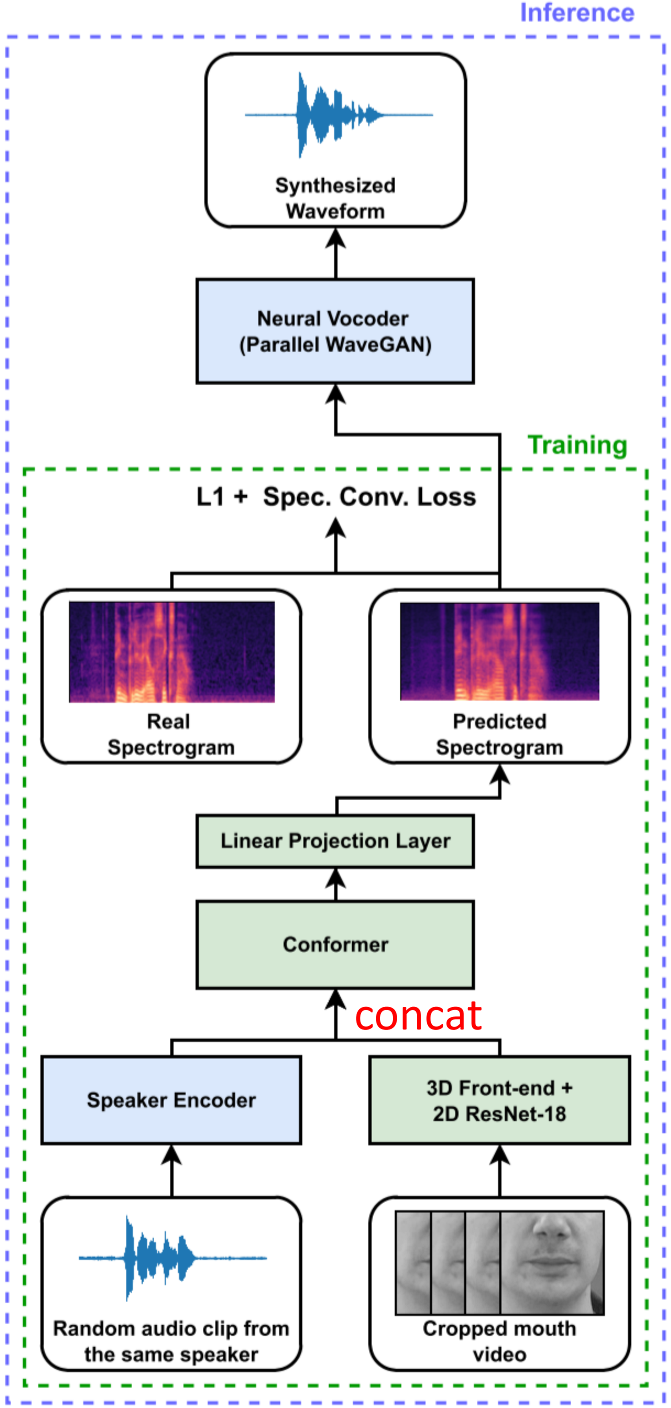
* Imperial College London, UK | University of Augsburg, Germany

- Methods
  - Video-to-spectrogram
    - 3D Front-end + 2D ResNet-18
    - Pre-trained speaker encoder
    - Conformer
    - L1 loss and the spectral convergence loss

$$L_{sc}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\| |\mathrm{STFT}(\boldsymbol{x})| - |\mathrm{STFT}(\hat{\boldsymbol{x}})| \|_F}{\| |\mathrm{STFT}(\boldsymbol{x})| \|_F},$$

  - Spectrogram-to-waveform
    - Pre-trained neural vocoder

Model	SVTS-S	SVTS-M	SVTS-L
Num. parameters* (M)	27.3	43.1	87.6
Conformer blocks	6	12	12
Attention dim.	256	256	512
Attention heads	4	4	8
Conv. kernel size	31	31	31
Feedforward dim.	2048	2048	2048



SVTS: Scalable Video-to-Speech Synthesis

demo:https://sites.google.com/view/scalable-vts

* arxiv 2022
* Imperial College London, UK | University of Augsburg, Germany

- Experiments

Table 2: *Summary of our results. Due to LRS3's complex vocabulary and long sentence structure, we are unable to find a speech recognition model that works accurately on our generated samples (e. g., the word "teacher" is often mistaken for "teachers"), and therefore do not report WER for this dataset.* **reported using Google speech-to-text API.*

Method	Corpus	Speaker split (seen/unseen)	Training data (hours)	PESQ	STOI	ESTOI	WER (%)
End-to-end GAN [24]	GRID	seen	24	1.70	0.667	0.466	4.60
VCA-GAN + Griffin-Lim [18]	GRID	seen	20	**1.97**	0.695	0.505	5.13
SVTS-S	GRID	seen	24	**1.97**	**0.705**	**0.523**	**2.36**
End-to-end GAN [38]	GRID	unseen	13	1.26	0.494	0.198	32.79
Conv. + GRU + WORLD vocoder [23]	GRID	unseen	13	1.26	0.541	0.227	38.15
End-to-end GAN [24]	GRID	unseen	13	1.37	0.568	0.289	**16.12**
VCA-GAN + Griffin-Lim [18]	GRID	unseen	13	1.39	0.570	0.282	24.57
Conv. + LSTM + WaveNet [16]	GRID	unseen	13	1.33	0.531	0.271	26.17
SVTS-S	GRID	unseen	13	**1.40**	**0.588**	**0.318**	17.85
Conv. + LSTM + Griffin-Lim [32]	LRW	unseen	157	1.20	0.543	0.344	34.20*
End-to-end GAN [24]	LRW	unseen	157	1.33	0.552	0.330	42.60
VCA-GAN + Griffin-Lim [18]	LRW	unseen	157	1.34	0.565	0.364	37.07
SVTS-M	LRW	unseen	157	**1.49**	**0.649**	**0.483**	**13.40**
SVTS-L	LRS3	seen	256	**1.30**	**0.553**	**0.331**	-
SVTS-L	LRS3	unseen	296	1.25	0.507	0.271	-
SVTS-L	LRS3 + VoxCeleb2	unseen	1556	**1.26**	**0.530**	**0.313**	-

WER on GT
GRID: 0.1%
LRW : 1.68%

SVTS: Scalable Video-to-Speech Synthesis

* arxiv 2022

* Imperial College London, UK | University of Augsburg, Germany

- Experiments : Ablations

Table 3: *Vocoder ablation on GRID (seen speakers). Speed is measured on an Nvidia RTX 2080 Ti. *computed on CPU*

Metric	PESQ	STOI	ESTOI	WER (%)	Speed (clips/sec.)
Griffin-Lim* [12]	**2.00**	0.696	0.513	2.41	1.2
Multiband MelGAN [41]	1.86	0.683	0.487	2.50	**184.9**
Style MelGAN [25]	1.93	0.702	0.520	2.38	83.7
Parallel WaveGAN [40]	1.97	**0.705**	**0.523**	**2.36**	54.7

Table 4: *Loss ablation on GRID (seen speakers).*

Metric	PESQ	STOI	ESTOI	WER (%)
w/o Spec. Conv.	**1.97**	**0.705**	**0.523**	2.90
w/o L_1	1.91	0.700	0.514	2.74
L_1+Spec. Conv.	**1.97**	**0.705**	**0.523**	**2.36**

【腾讯文档】VTS
https://docs.qq.com/sheet/DYVFlblV1dWNuZ0R2