Audio-Visual Speech Codecs:
Rethinking Audio-Visual Speech Enhancement by Re-Synthesis

Karren Yang[1] Dejan Marković[2] Steven Krenn[2] Vasu Agrawal[2] Alexander Richard[2]
[1]MIT [2]Meta Reality Labs Research
karren@mit.edu {dejanmarkovic,stevenkrenn,vasuagrawal,richardalex}@fb.com

CVPR 2022, https://arxiv.org/pdf/2203.17263.pdf

Introduction

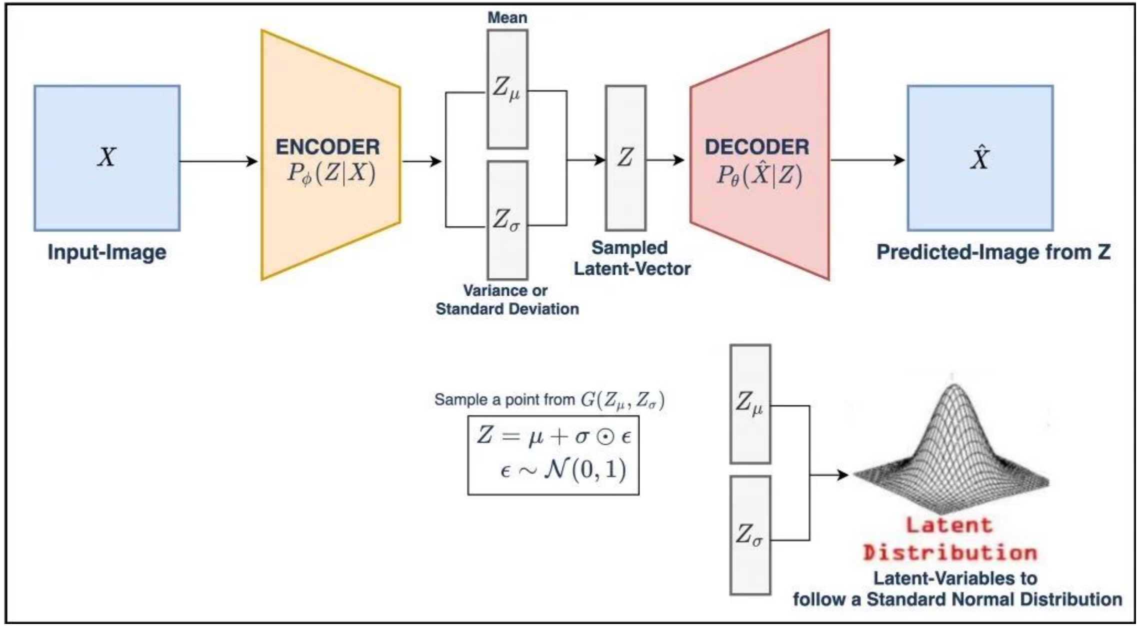
- Denoising Speech Signal

$$x(t) + n(t) = y(t)$$

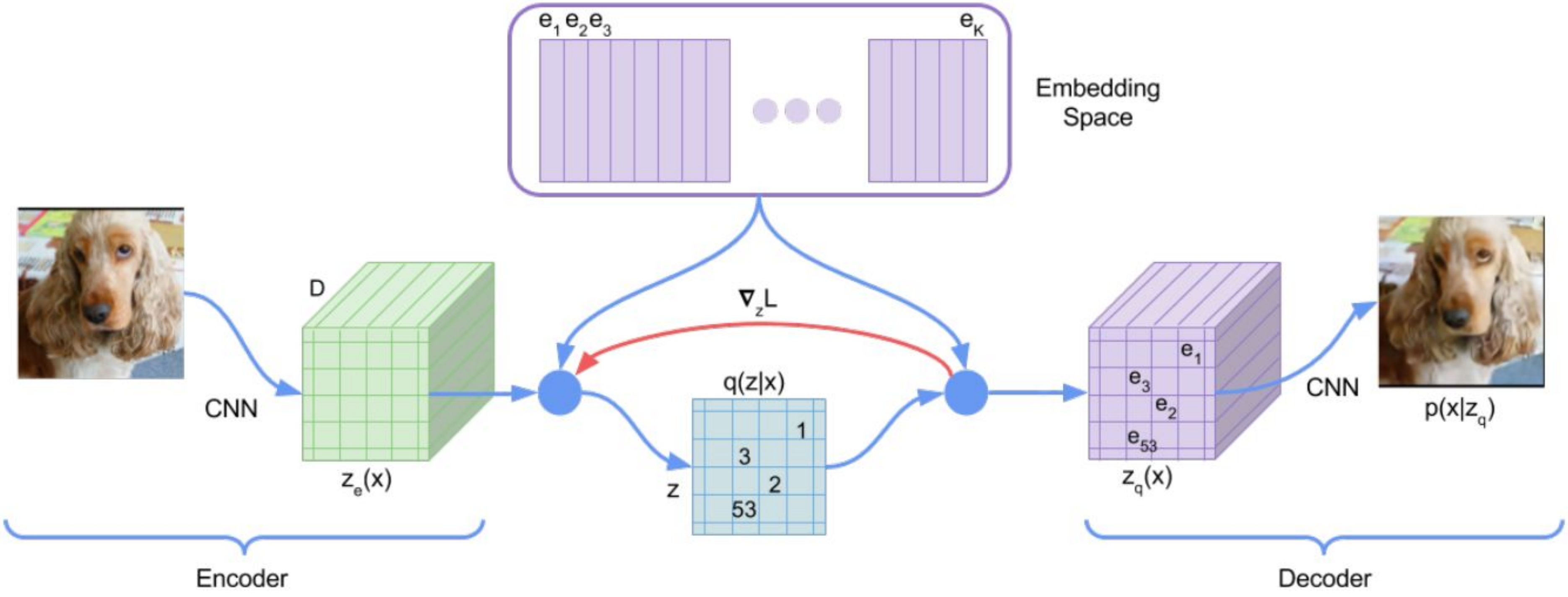
- Using Visual Information as auxiliary signal
- Used Generative approach (VAE)

Approach

- Used Vector Quantized VAE (VQ-VAE)
- Discrete Representation



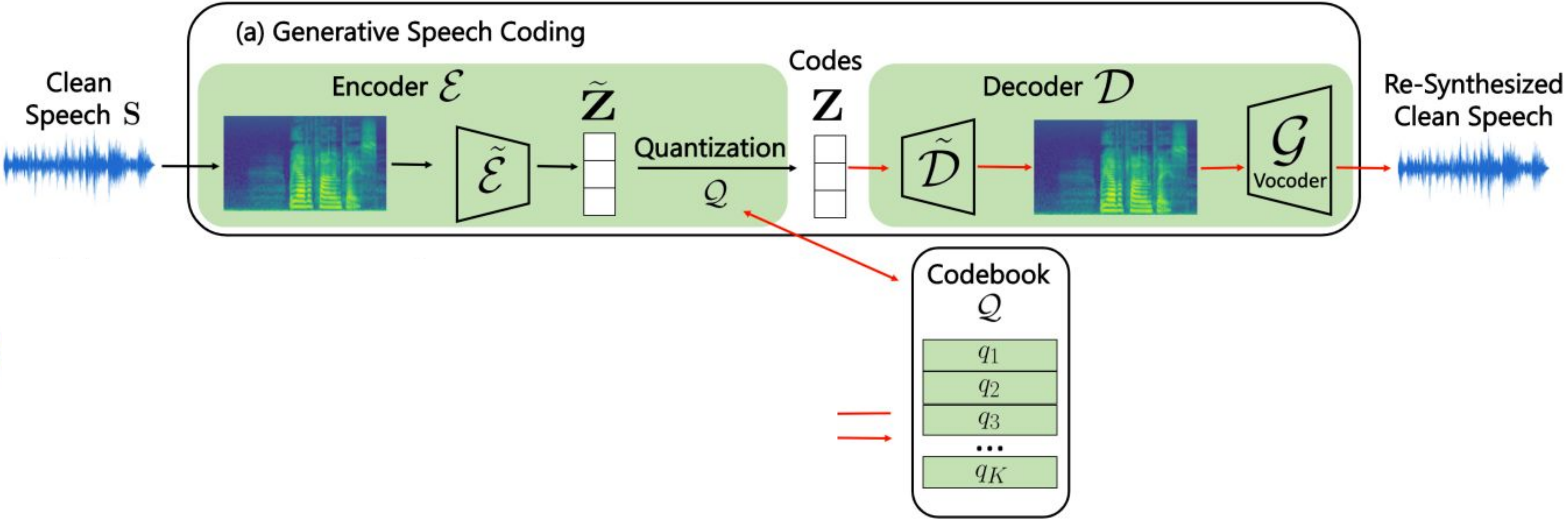
VQ-VAE



e1 e2 e3
eK
Embedding Space
D
CNN
ze(x)
∇zL
q(z|x)
1
3
2
53
z
e1
e3
e2
e53
zq(x)
CNN
p(x|zq)
Encoder
Decoder

Approach

$$\mathbb{E}_{(\mathbf{S},\mathbf{S}',\mathbf{V})}||\mathbf{melspec}(\mathbf{S}) - \tilde{\mathcal{D}}(\mathbf{Z})||_2^2$$



Dataset

FaceStar

- Audio-visual dataset containing 10 hours of speech data from two speakers

AV Dataset	# Hours per Speaker	High-Quality Audio	Reliable Lip Motion	Unconstrained Natural Speech
GRID	0.8	✓	✓	✗
TCD-Timit	0.5	✓	✓	✗
Lip2Wav	20	✗	✗	✓
Facestar	5	✓	✓	✓

Results

Model	Facestar PESQ ↑	Facestar STOI ↑	Facestar F-SNR ↑	Facestar MCD ↓	Facestar Mel-ℓ_2 ↓	Lip2Wav PESQ ↑	Lip2Wav STOI ↑	Lip2Wav F-SNR ↑	Lip2Wav MCD ↓	Lip2Wav Mel-ℓ_2 ↓
Demucs [8]	1.251	0.554	5.602	5.003	0.0106	1.383	0.672	7.644	4.724	0.0109
AV-Masking [18]	1.257	0.593	5.991	5.184	0.0093	1.438	0.689	7.873	5.167	0.0093
AV-Mapping [15]	1.332	0.626	2.802	4.885	0.0059	1.417	0.661	6.892	4.643	0.0062
Ours	**1.354**	**0.661**	**7.322**	**3.815**	**0.0056**	**1.482**	**0.740**	**8.801**	**4.072**	**0.0055**

Ours	GT recordings	Can not tell
4.1%	44.5%	51.4%

Ours	AV Encoder Decoder	Can not tell
73.3%	6.0%	20.7%

Ours	AV Masking	Can not tell
78.5%	5.7%	15.8%

Table 3. **Perceptual Evaluation**. Participants were presented two video clips and asked to tell which of the two sounds more natural.

Model	reverb + noise + interfering spkr	only reverb + noise
Vision-Only	0.0085	
Audio-Only	0.0091	0.0056
No Auto-Regressive Module	0.0051	0.0036
Full Model	**0.0043**	**0.0033**

Table 4. **Ablation Results**. The values shown are the mean ℓ_2 errors between predicted and ground truth mel-spectrograms for ablation models trained on the Facestar dataset (Speaker 1); lower is better. See text for details.

Multi-Speaker Model

- Extend framework to multispeaker setting by adding a speaker identity encoder

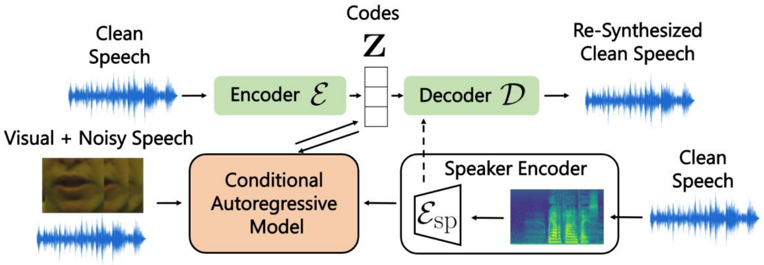


Figure 4. **Multi-speaker model**. A speaker encoder is added to the pipeline from Figure 2. Restricting the size of the codebook forces the model to disentangle speech content and speaker identity as shown in [48].

	GRID Speaker			
	Sp. 1 (M)	**Sp. 3 (M)**	**Sp. 11 (F)**	**Sp. 15 (F)**
Single-speaker model	0.00509	0.00794	0.00746	0.00781
Multi-speaker model	0.00657	0.00909	0.00960	0.01594
Multi-speaker model personalized to new speaker with k minutes of data				
5 min	0.00481	0.00682	0.00625	0.00681
12.5 min	0.00457	0.00620	0.00589	0.00655
25 min	0.00443	0.00595	0.00570	0.00621
50 min	0.00425	0.00561	0.00553	0.00596

Voice Controllability

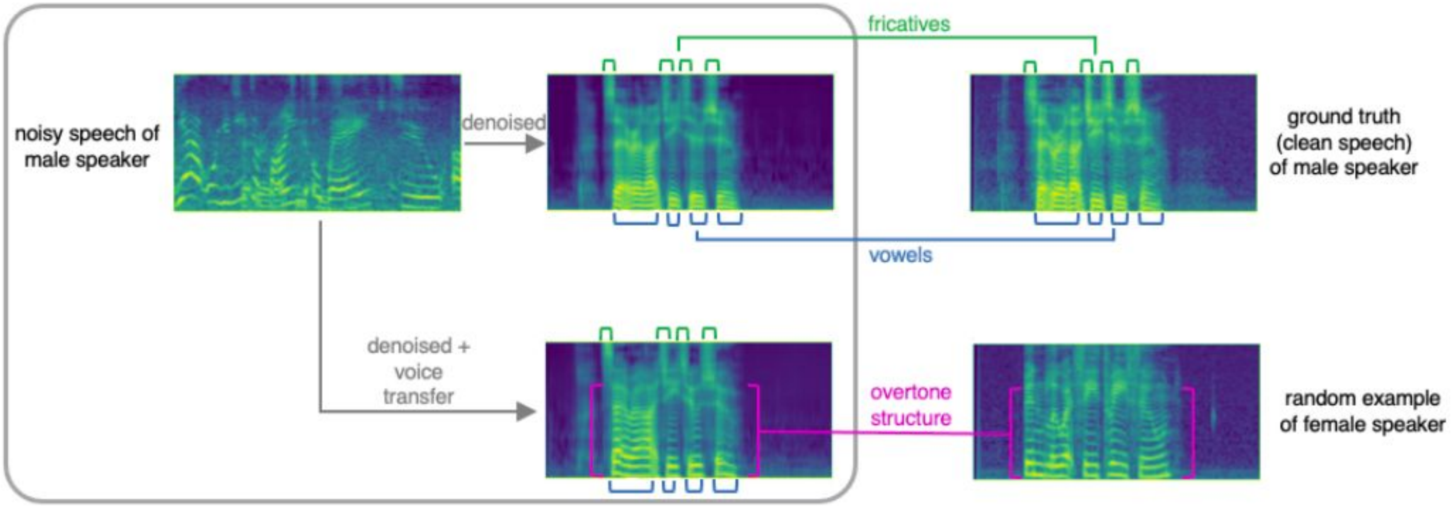


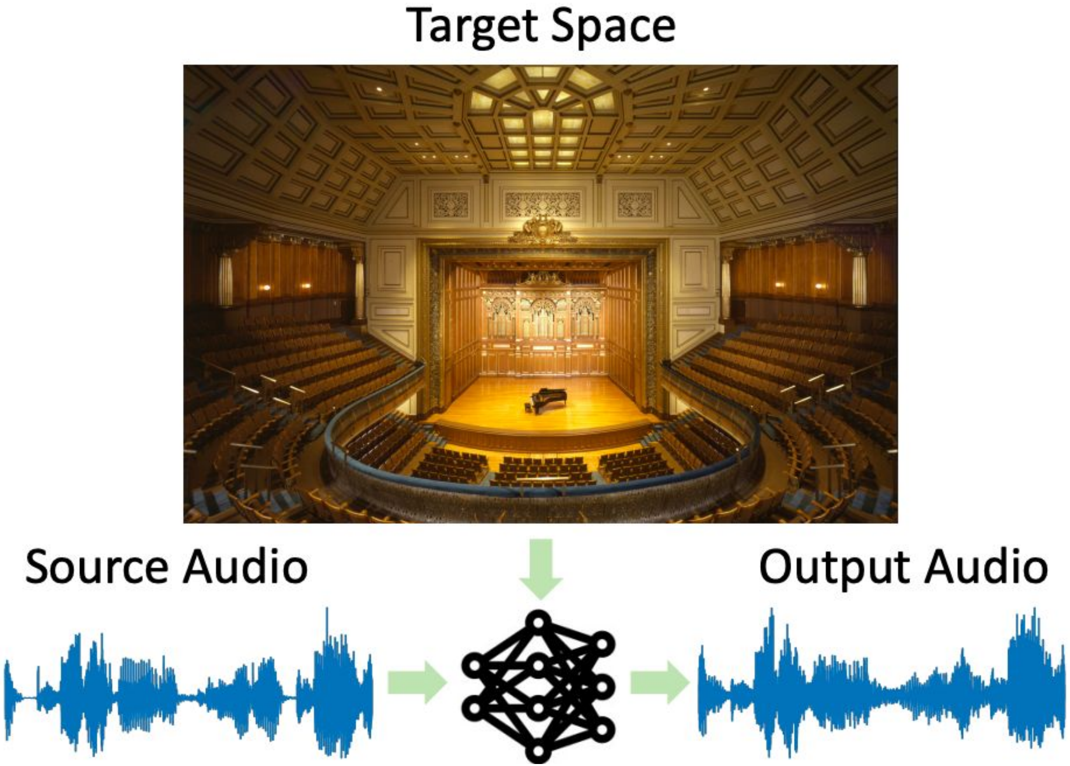
Figure 5. **Voice Transfer Examples.** By swapping the speaker code at the decoder stage, we can synthesize clean audio in a different target speaker's voice. Images shown are mel-spectrogram representations of audio. Note how the linguistic content (*i.e.*, vowels and fricatives) are carried over from the original male speaker, while the pitch and overtone structure are changed to that of the female speaker.

Visual Acoustic Matching

Changan Chen[1,4] Ruohan Gao[2] Paul Calamia[3] Kristen Grauman[1,4]
[1]University of Texas at Austin [2]Stanford University [3]Facebook Reality Labs [4]Facebook AI Research

CVPR 2022, https://arxiv.org/pdf/2202.06875.pdf

Introduction



$$A_r(t) = A_s(t) * R(t)$$

Target Audio

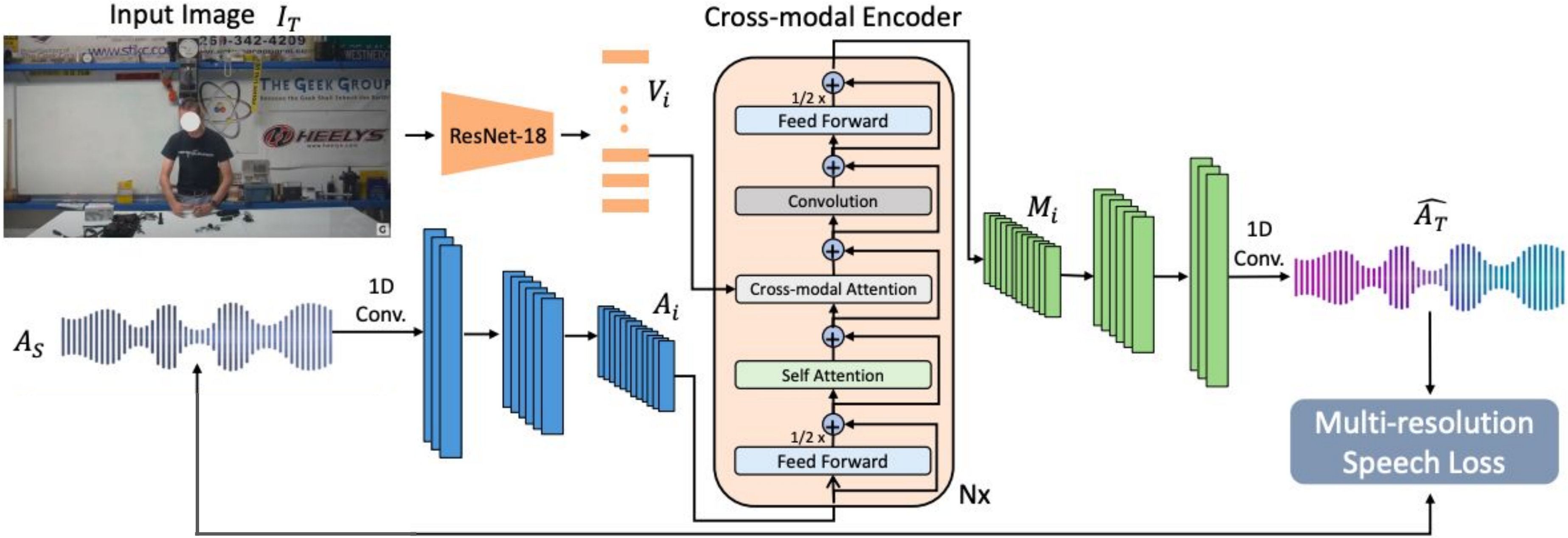
Source Audio

RIR

Challenges

- Unpaired data
- Modelling different regions of the room
- Capture geometry and materials present in the scene

Approach

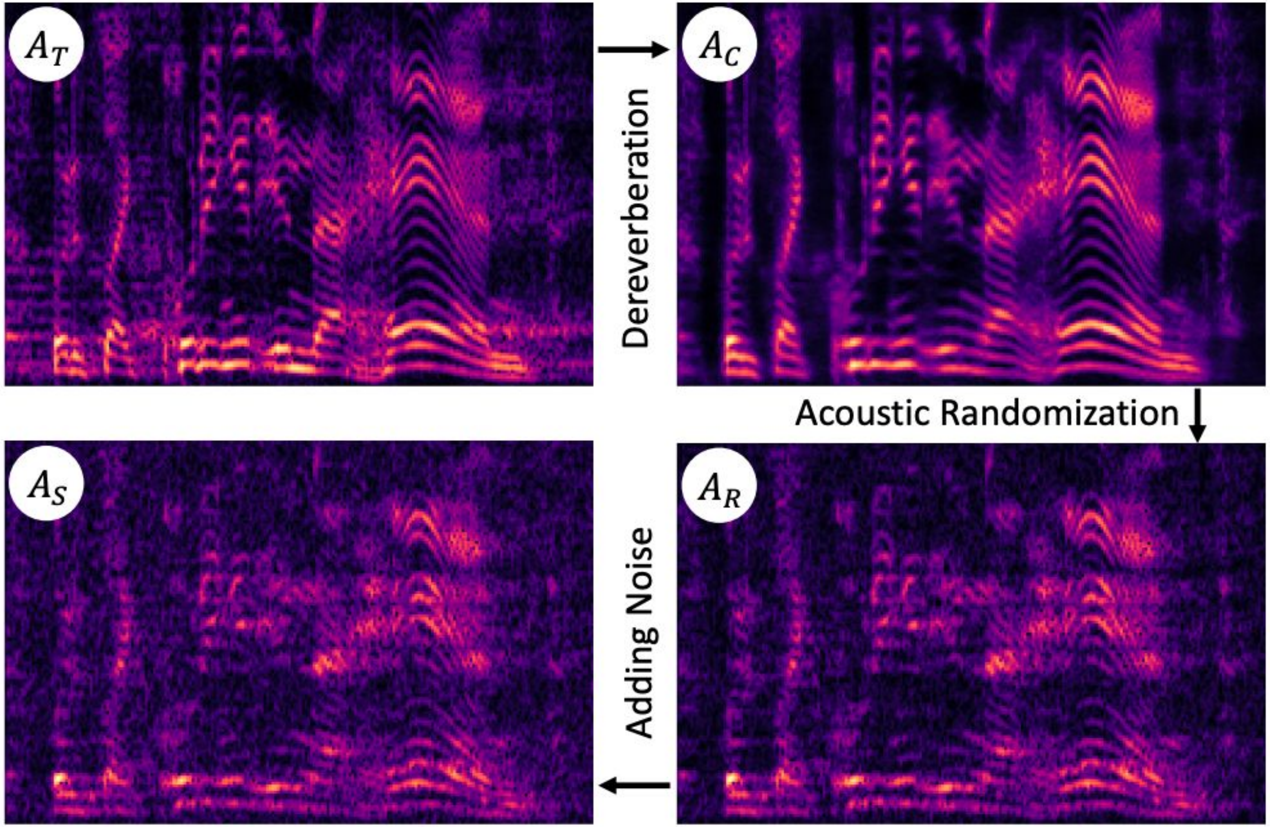


Input Image I_T
ResNet-18
V_i
Cross-modal Encoder
1/2 x
Feed Forward
Convolution
Cross-modal Attention
Self Attention
1/2 x
Feed Forward
Nx
A_S
1D
Conv.
A_i
M_i
1D
Conv.
$\widehat{A_T}$
Multi-resolution
Speech Loss

Datasets

1. SoundSpaces-Speech Dataset
    - Synthetic Data on Matterport3D
    - RIR convolved with audio from LibriSpeech
    - 3D humanoid inserted at speaker location
2. AVSpeech
    - Subset of AVSpeech dataset
    - 3-10 seconds videos with single visible human speaker

Acoustic Alteration Process



Loss Function

$$\mathcal{L}_G = \sum_{k=1}^{K} \left(\mathcal{L}_{Adv}(G; D_k) + \lambda_1 \mathcal{L}_{FM}(G; D_k)\right) + \lambda_2 \mathcal{L}_{Mel}(G)$$

$$\mathcal{L}_D = \sum_{k=1}^{K} \mathcal{L}_{Adv}(D_k; G)$$

Evaluation Metric

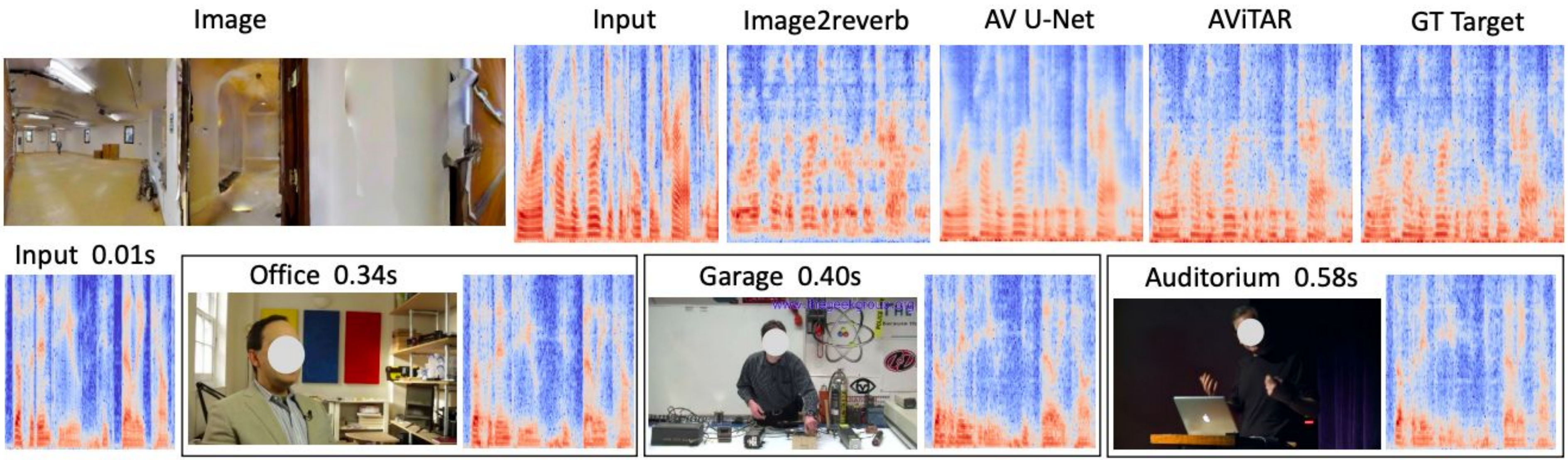
- STFT Distance
- RT60 Error (RTE)
- Mean Opinion Score Error (MOSE)
- User Study

Results

	SoundSpaces-Speech						*Acoustic AVSpeech*			
	Seen			*Unseen*			*Seen*		*Unseen*	
	STFT	RTE (s)	MOSE	STFT	RTE (s)	MOSE	RTE (s)	MOSE	RTE (s)	MOSE
Input audio	1.192	0.331	0.617	1.206	0.356	0.611	0.387	0.658	0.392	0.634
Blind Reverberator [64]	1.338	0.044	0.312	-	-	-	-	-	-	-
Image2Reverb [55]	2.538	0.293	0.508	2.318	0.317	0.518	-	-	-	-
AV U-Net [22]	**0.638**	0.095	0.353	**0.658**	0.118	0.367	0.156	0.570	0.188	0.540
AViTAR w/o visual	0.862	0.140	0.217	0.902	0.186	0.236	0.194	0.504	0.207	0.478
AViTAR	0.665	**0.034**	**0.161**	0.822	**0.062**	**0.195**	**0.144**	**0.481**	**0.183**	**0.453**

	SoundSpaces	AVSpeech
Input Speech	42.1% / **57.9%**	40.1% / **59.9%**
Image2Reverb [55]	25.9% / **74.1%**	- / -
AV U-Net [22]	29.8% / **70.2%**	27.2% / **72.8%**
AViTAR w/o visual	39.6% / **60.4%**	46.3% / **53.9%**

Qualitative Results



https://vision.cs.utexas.edu/projects/visual-acoustic-matching/

Self-supervised object detection from audio-visual correspondence

Triantafyllos Afouras[1*†] Yuki M. Asano[1*] Francois Fagan[2] Andrea Vedaldi[2] Florian Metze[2]
[1] Visual Geometry Group, University of Oxford
[2] Facebook AI
afourast@robots.ox.ac.uk

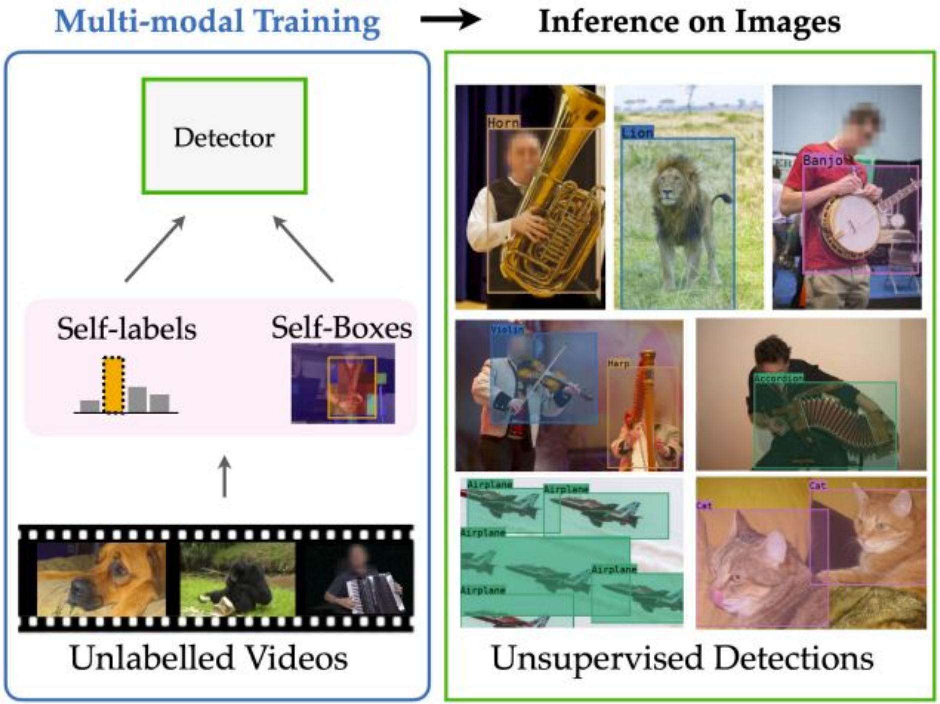
CVPR 2022, https://arxiv.org/pdf/2104.06401.pdf

Object Detection

- Supervised
- Weakly-Supervised
- Unsupervised

Introduction

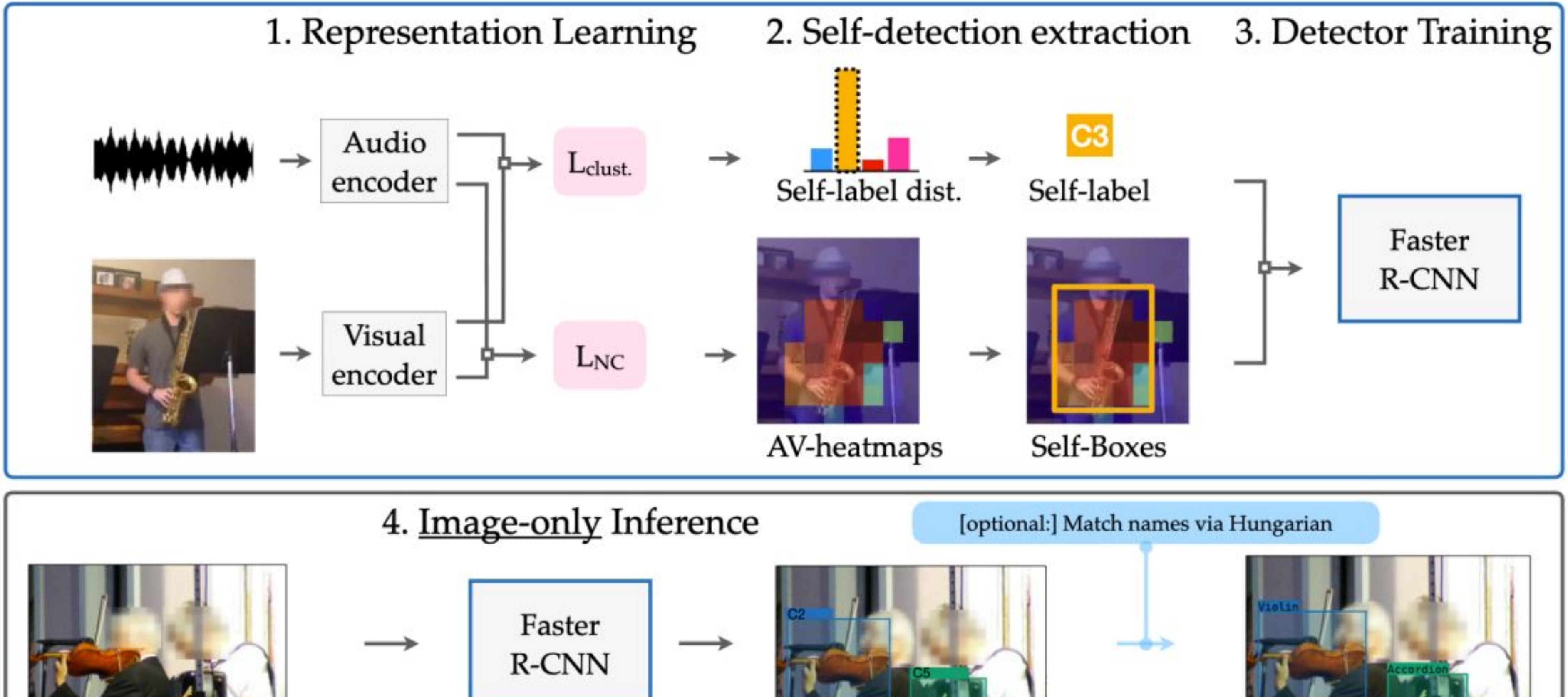
- Object Detection using Audio



Problems with audio-visual detection

- Not applicable to silent videos
- Only heatmap, No bounding box

Approach



1. Representation Learning
2. Self-detection extraction
3. Detector Training
Audio encoder
Visual encoder
$L_{clust.}$
L_{NC}
Self-label dist.
C3
Self-label
AV-heatmaps
Self-Boxes
Faster R-CNN
4. Image-only Inference
[optional:] Match names via Hungarian
Faster R-CNN
C2
C5
Violin
Accordion

Results

Method	single-instr. IoU-0.5	single-instr. AUC	multi-instr. cIoU-0.3
Sound of pixels [109]	38.2	40.6	39.8
Object t. Sound [7]	32.7	39.5	27.1
Attention [82]	36.5	39.5	29.9
DMC [44]	32.8	38.2	32.0
DSOL [45]	38.9	40.9	48.7
Ours	**50.6**	**47.5**	**52.4**

		VGGSound			Audioset			OpenImages		
Method	No labels?	mAP_{30}	mAP_{50}	$mAP_{[50:95:5]}$	mAP_{30}	mAP_{50}	$mAP_{[50:95:5]}$	mAP_{30}	mAP_{50}	$mAP_{[50:95:5]}$
PCL (WSOD) [90]	✗	54.9	27.7	7.6	39.0	17.5	4.4	37.9	14.5	3.5
Ours - weak sup.	✗	67.6	42.9	14.2	50.6	30.9	10.3	48.9	33.7	9.5
Center Box*	✓	29.6	5.6	1.5	15.1	3.5	0.7	20.7	4.2	0.8
Selective Search* [94]	✓	5.2	1.1	0.4	2.8	0.4	0.1	7.4	2.1	0.7
COCO-trained RPN*	✗	33.4	7.5	1.6	19.0	4.1	0.8	24.4	11.1	2.6
Ours - self-boxes*	✓	48.1	29.6	10.0	27.8	14.1	4.8	NA	NA	NA
Ours - full	✓	**52.3**	**39.4**	**14.7**	**44.3**	**28.0**	**9.6**	**39.9**	**28.5**	**7.6**