

# Supplementary Material: Audio-Visual Speech Super-Resolution

Rudrabha Mukhopadhyay\*<sup>1</sup>

radrabha.m@research.iiit.ac.in

Sindhu B Hegde\*<sup>1</sup>

sindhu.hegde@research.iiit.ac.in

Vinay Namboodiri<sup>2</sup>

vpn22@bath.ac.uk

C.V. Jawahar<sup>1</sup>

jawahar@iiit.ac.in

<sup>1</sup> IIIT-Hyderabad,

Gachibowli, Hyderabad  
India

<sup>2</sup> University of Bath,

Claverton Down,  
Bath,  
United Kingdom

In this supplementary document, we report the architectural details of our network and various additional experiments that we conducted. We deem these experiments to be insightful and help us to better understand different components of our network.

## 1 Model Architecture

In addition to the description provided in Section 3 of the main paper, here, we detail the architecture of the different modules of our proposed audio-visual speech super-resolution (SR) network.

### 1.1 Speech Encoder

We convert the raw speech into the linear spectrograms using STFT (as described in Section 3.1 in the paper) and concatenate the magnitude and the phase components to obtain  $T \times 514$  dimensions representation. Our speech encoder ingests this representation and processes them with a stack of 1D convolution layers as described in Table 1 to generate speech embeddings (dimension:  $T \times 600$ ).

### 1.2 Visual Encoder

The visual encoder ingests the input frame sequences of dimension  $T/4 \times 3 \times 96 \times 96$ . It generates visual embeddings of dimension  $T \times 600$  using 3D convolution layers as described in Table 2.

Table 1: Details of the speech encoder.

Layer	# Filters	Kernel	Stride	Residual	Output
input	-	-	-	-	$T \times 257$
conv1	600	3	1	$\times$	$T \times 600$
conv2	600	3	1	$\checkmark$	$T \times 600$
conv3	600	3	1	$\checkmark$	$T \times 600$
conv4	600	3	1	$\checkmark$	$T \times 600$
conv5	600	3	1	$\checkmark$	$T \times 600$
conv6	600	3	1	$\checkmark$	$T \times 600$
conv7	600	3	1	$\times$	$T \times 600$

Table 2: Details of the visual encoder.

Layer	# Filters	Kernel	Stride	Residual	Output
input	3	-	-	-	$T/4 \times 3 \times 96 \times 96$
transpose1	-	-	-	-	$3 \times T/4 \times 96 \times 96$
conv1	32	$5 \times 5 \times 5$	$1 \times 2 \times 2$	$\times$	$32 \times T/4 \times 48 \times 48$
conv2	32	$5 \times 5 \times 5$	$1 \times 1 \times 1$	$\checkmark$	$32 \times T/4 \times 48 \times 48$
conv3	64	$1 \times 3 \times 3$	$1 \times 2 \times 2$	$\times$	$64 \times T/4 \times 24 \times 24$
conv4	64	$1 \times 3 \times 3$	$1 \times 1 \times 1$	$\checkmark$	$64 \times T/4 \times 24 \times 24$
conv5	128	$1 \times 3 \times 3$	$1 \times 2 \times 2$	$\times$	$128 \times T/4 \times 12 \times 12$
conv6	128	$1 \times 3 \times 3$	$1 \times 1 \times 1$	$\checkmark$	$128 \times T/4 \times 12 \times 12$
conv7	256	$1 \times 3 \times 3$	$1 \times 2 \times 2$	$\times$	$256 \times T/4 \times 6 \times 6$
conv8	256	$1 \times 3 \times 3$	$1 \times 1 \times 1$	$\checkmark$	$256 \times T/4 \times 6 \times 6$
conv9	512	$1 \times 3 \times 3$	$1 \times 2 \times 2$	$\times$	$512 \times T/4 \times 3 \times 3$
conv10	512	$1 \times 3 \times 3$	$1 \times 1 \times 1$	$\checkmark$	$512 \times T/4 \times 3 \times 3$
conv11	600	$1 \times 3 \times 3$	$1 \times 3 \times 3$	$\times$	$600 \times T/4 \times 1 \times 1$
conv12	600	$1 \times 1 \times 1$	$1 \times 1 \times 1$	$\times$	$600 \times T/4 \times 1 \times 1$
transpose2	-	-	-	-	$T/4 \times 600 \times 1 \times 1$
squeeze	-	-	-	-	$T/4 \times 600$
upsample	-	-	-	-	$T \times 600$

### 1.3 Speech Decoder

The speech and the visual embeddings obtained are concatenated to form a fused representation of dimension  $T \times 1200$  which is given as input to the speech decoder. The decoder comprises a series of 1D convolution layers and aims to generate residual mask of dimension  $T \times 514$  as shown in Table 3.

## 2 Additional Experiments

### 2.1 Quantitative Evaluation

In addition to the qualitative evaluation on VoxCeleb2 dataset [9], we further assess our model on the official test set of LRS2 dataset [2]. Note that we do not fine-tune our model on the LRS2 [2] dataset; thus, evaluating on it demonstrates the generalisation ability of our model on new datasets (with significantly different pre-processing and image resolutions used during data collection).

Table 3: Details of the speech decoder.

Layer	# Filters	Kernel	Stride	Residual	Output
input	-	-	-	-	$T \times 1200$
conv1	1024	3	1	$\times$	$T \times 1024$
conv2	1024	3	1	$\checkmark$	$T \times 1024$
...					
conv13	1024	3	1	$\checkmark$	$T \times 1024$
conv14	1024	3	1	$\checkmark$	$T \times 1024$
conv15	514	3	1	$\times$	$T \times 514$

Table 4: Quantitative comparison of different approaches at scale factors of  $8\times$  and  $16\times$  on LRS2 [10] dataset.

Scale factor	$8\times$				$16\times$			
Method	PESQ $\uparrow$	STOI $\uparrow$	ESTOI $\uparrow$	LSD $\downarrow$	PESQ $\uparrow$	STOI $\uparrow$	ESTOI $\uparrow$	LSD $\downarrow$
Linear	2.201	0.753	0.558	10.771	2.017	0.570	0.318	11.612
TFiLM [10]	2.191	0.768	0.675	7.103	2.042	0.679	0.491	8.623
NU-Wav [10]	2.250	0.761	0.651	7.946	2.005	0.628	0.524	8.711
AO baseline	1.914	0.802	0.692	6.242	1.706	0.701	0.525	8.007
<b>Ours (pseudo)</b>	2.584	0.808	0.702	6.005	2.616	0.739	0.622	6.991
<b>Ours</b>	<b>2.805</b>	<b>0.815</b>	<b>0.725</b>	<b>5.197</b>	<b>2.637</b>	<b>0.766</b>	<b>0.649</b>	<b>5.838</b>

In-line with the results on VoxCeleb2 dataset [10] as shown in Section 4.2.1 in the main paper, our model performs remarkably well compared to the existing audio-only approaches as shown in Table 4 on LRS2 dataset [10]. At both the scale factors, our model performs consistently better, thereby significantly improving the generated speech quality and intelligibility. Also, our pseudo-visual model achieves substantial boost in comparison to all the audio-only approaches, although neither the pseudo-visual nor the speech SR models have been fine-tuned on LRS2 data [10]. This demonstrates the robustness and generalisation ability of our method to different identities and datasets.

## 2.2 Model’s Variation to Identity Attributes

We analyse the behaviour of our speech SR model on the identity attributes such as gender and age (from the test set of VoxCeleb2 data [10]) in Table 5. For gender classification, we use a gender detection tool [10] which automatically groups the identities into male and female categories. To identify the age of the speakers, we use the public implementation from<sup>1</sup>. We can see from Table 5 that our speech SR network is consistent across the different age groups, but scores do vary slightly across the gender of the identities.

Table 5: Effect of the identity attributes such as gender and age on model’s performance.

Attribute	Class	PESQ $\uparrow$	STOI $\uparrow$	ESTOI $\uparrow$	LSD $\downarrow$
Gender	Female	2.562	0.740	0.687	5.515
	Male	2.520	0.784	0.636	5.504
Age	< 25	2.213	0.776	0.665	5.652
	25 – 50	2.287	0.751	0.650	5.727
	> 50	2.271	0.773	0.638	5.452

<sup>1</sup><https://github.com/yu4u/age-estimation-pytorch>

## 2.3 Comparison of Pseudo-Lip Identities

We compare the performance of our pseudo-visual model when different static identities are used for the generation of lip movements. Table 6 shows the results on the test set of VoxCeleb2 data [9]. We can see that the model’s performance does not vary across the pseudo-lip identities.

Table 6: Our pseudo-visual model is invariant to pseudo-lip identities.

Identities	PESQ↑	STOI↑	ESTOI↑	LSD↓
Taylor Swift	2.237	0.762	0.651	5.500
Paul McCartney	2.218	0.763	0.643	5.618
Barack Obama	2.239	0.758	0.645	5.498

## References

- [1] Sawyer Birnbaum, Volodymyr Kuleshov, Zayd Enam, Pang Wei W Koh, and Stefano Ermon. Temporal film: Capturing long-range sequence dependencies with feature-wise modulations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [2] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [3] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.
- [4] Junhyeok Lee and Seungu Han. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *INTERSPEECH*, 2021.
- [5] Arun Ponnusamy. cvlib - a high level computer vision library for python. <https://github.com/arunponnusamy/cvlib>, 2018.