Audio-Visual Speech Super-Resolution

Rudrabha Mukhopadhyay[*1]
radrabha.m@research.iiit.ac.in

Sindhu B Hegde[*1]
sindhu.hegde@research.iiit.ac.in

Vinay Namboodiri[2]
vpn22@bath.ac.uk

C.V. Jawahar[1]
jawahar@iiit.ac.in

[1] IIIT-Hyderabad,
Gachibowli, Hyderabad
India

[2] University of Bath,
Claverton Down,
Bath,
United Kingdom

Abstract

In this paper, we present an audio-visual model to perform speech super-resolution at large scale-factors (8× and 16×). Previous works attempted to solve this problem using only the audio modality as input, and thus were limited to low scale-factors of 2× and 4×. In contrast, we propose to incorporate both visual and auditory signals to super-resolve speech of sampling rates as low as 1kHz. In such challenging situations, the visual features assist in learning the content, and improves the quality of the generated speech. Further, we demonstrate the applicability of our approach to arbitrary speech signals where the visual stream is not accessible. Our "pseudo-visual network" precisely synthesizes the visual stream solely from the low-resolution speech input. Extensive experiments illustrate our method's remarkable results and benefits over state-of-the-art audio-only speech super-resolution approaches. Our project website can be found at `http://cvit.iiit.ac.in/research/projects/cvit-projects/audio-visual-speech-super-resolution`.

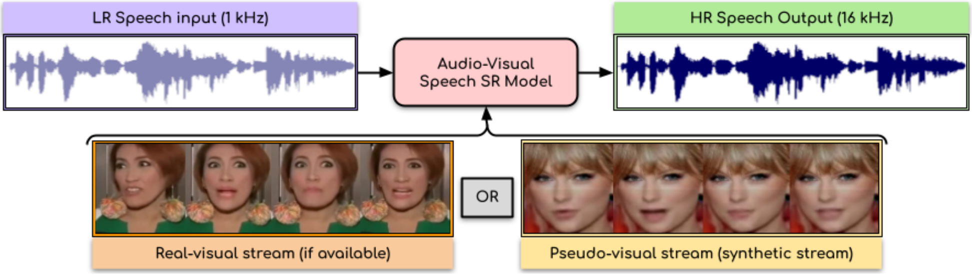
LR Speech input (1 kHz)
Audio-Visual Speech SR Model
HR Speech Output (16 kHz)
OR
Real-visual stream (if available)
Pseudo-visual stream (synthetic stream)

Figure 1: We present an audio-visual model for super-resolving very low-resolution speech inputs (example, 1kHz) at large scale-factors. In contrast to the existing audio-only speech super-resolution approaches, our method benefits from the visual stream, either the real-visual stream (if available), or the generated visual stream from our pseudo-visual network.

1 Introduction

Conversations in our everyday life are constantly being corrupted by degradation to speech signals (for example, electronic transmission, and background noise). This has attracted several works with profound interests to recover the useful signal from the corrupted mixture.

[*] Both the authors have contributed equally to this project.

Tasks such as speech denoising [14, 26, 31], speaker separation [1, 12], and noise-robust automatic speech recognition [17, 21, 23] have gained considerable progress in reconstructing high-quality speech from corrupted inputs. This work aims to handle a different form of degradation to the speech input: "low sampling rate". Learning a network to super-resolve speech would enable a wide variety of real-world applications. Some of them include: (i) recovering historical public talks and speeches, (ii) enhancing inaudible voices in the videos, (iii) improving the telephonic user experience by upsampling and rendering high-quality speech, and (iv) compressing the speech to reduce the bandwidth consumption. Further, several downstream tasks like automatic speech recognition [4] and speaker identification [10] could greatly benefit from speech super-resolution (SR).

Recovering the high-frequency information when the input sampling rate is very low (for example, 1kHz) is a significant challenge. When the input resolution is so low, the crucial information present in the speech, including the content and the voice attributes such as prosody, pitch, and style, are almost entirely lost. Thus, although speech SR has been extensively studied in the audio processing literature [5, 18, 20], most of these approaches lead to sub-optimal results for such low-resolution (LR) inputs. In addition, as these methods are designed for low scale-factors of 2× and 4×, they are not directly extensible for higher factors, severely limiting their real-world applicability. This can be mainly attributed to the inadequate prior information considered in the existing works to solve this task, i.e., only the audio modality. While there have been substantial improvements by incorporating the visual cues in multiple speech generation tasks [1, 12], speech SR has not yet witnessed the benefits from the visual stream. This raises an interesting question, "Can we use the visual stream to super-resolve very low-resolution speech?".

In this work, we aim to upsample very low-resolution speech signals (of sampling rate 1kHz) by proposing an "audio-visual" network. Inspired by the recent success of visual assistance in tasks such as speech enhancement [1, 12, 14, 24], and speech recognition [2], we propose to incorporate the visual modality for speech SR for the first time. We hypothesize that using the visual assistance is quintessential to recover the content from very low-resolution inputs, and significantly improve the speech quality, clarity, intelligibility, and naturalness.

We demonstrate the applicability of our approach for arbitrary in-the-wild speech signals, which need not necessarily have an associated visual stream. In such cases, we propose a "pseudo-visual model" to synthesize the visual stream from the LR speech input. We design a student-teacher training setup to synthesize the lip movements, which are further utilized as the input visual stream in our speech SR network. We are the first to perform 16× speech SR from an input speech with sampling rate as low as 1kHz. We demonstrate that our approach performs significantly better than the state-of-the-art audio-only approaches in all speech quality, and intelligibility metrics through extensive experimentation.

To summarize, our significant contributions are: (i) We solve the problem of speech SR at extreme scale-factors of 8× and 16× which was previously never accomplished in audio processing literature; (ii) We propose an audio-visual speech SR model that outperforms the audio-only models, and is applicable in real-world, unconstrained settings; (iii) We show that our model can be utilized even when the real-visual stream is unreliable or inaccessible. Our pseudo-visual network solely considers the LR speech, and synthesizes accurate lip movements, which are then ingested as the visual stream input by our speech SR network.

We provide a demo video in our project website at http://cvit.iiit.ac.in/research/projects/cvit-projects/audio-visual-speech-super-resolution, which exhibits the dominance of our approach compared to the audio-only works. The code and the trained models are released publicly to encourage future research. In Section 2, we review the existing works in this space. We then discuss the proposed audio-visual speech SR network in Section 3, and present the experimental results in Section 4. This is followed by an analysis of the different modules used in our architecture in Section 5, and finally, we conclude our work in Section 6.

2 Related Work

Audio-only Speech Super-Resolution: The speech community has studied the problem of upsampling the frequency of speech signals for a long time. This problem was popularly known as "bandwidth extension" in pre-deep learning era. Initially, classical signal processing approaches were used [11, 19] to solve the task of bandwidth extension. This was followed by methods depending on the Gaussian mixture models to predict the high-frequency speech based solely on the low-frequency input [6]. Soon, deep learning led to renewed interest in this problem under a new alias, i.e., "audio super-resolution". Inspired by the success of image super-resolution techniques using deep learning, Li et al. [22] were the first to use a simple neural network to learn mappings between high-resolution and low-resolution audio signals. This was further improved by various techniques like residual-based bottleneck network [18], "Temporal-Film (TFiLM)" [5], and diffusion probabilistic model "NU-Wav" [20], significantly enhancing the SR performance. While the current state-of-the-art methods work directly on low-resolution speech, they are proposed for 2× and 4× SR which is in stark contrast to our attempt of 16× SR.

To improve these networks' robustness and real-world applicability, we consider using additional assistance in terms of the visual modality, particularly the lip movements. In recent years, several impressive works have been proposed for other related problems where the visual stream is used to boost the quality of the generated speech. We now review these advancements in the multi-modal space, and discuss how we incorporate visual assistance for the task at hand.

Correlation between Speech and Lip Movements: Cross-modal assistance for audio and visual modalities has proven beneficial in various tasks such as object localization [3, 25], audio source separation, and denoising [1, 12, 14], image animation [7, 28], audio-visual speech recognition [2], and speech generation from lip movements [27]. Since lip movements and speech are naturally correlated, the phoneme-viseme mapping between them is widely explored in recent times. Popular works like "The Conversation" [1] to isolate the individual speakers, a multi-stream network for solving the "cocktail-party" problem [12], and multisensory feature based model [24] to solve tasks such as sound localization, action recognition, and audio source separation, have shown remarkable results by using the visual cues. These works attempted to train models by feeding "multi-speaker speech" and lip movements of the target speaker to isolate the speech corresponding to individual speakers. The idea of exploiting the lip movements to provide additional information about the clean speech was also extensively used in [14]. Here, the authors propose to generate a synthetic stream of information in the form of lip movements by considering the noisy speech as input. The generated pseudo-visual stream acts as a "visual noise filter" and helps to distill the clean speech from a given noisy speech segment.

Motivated by these advancements, we explore super-resolving low-resolution speech by using lip movements as additional cues. The assistance from the visual stream allows us to handle large scale-factors like 16× compared to 4× as done in the previous works. Further, to enable our model to be applied in practical situations, we extend the pseudo-visual approach proposed in [14] for speech SR. Thus, along with the audio-visual approach, we also develop an audio-only system that incorporates the advantages of the visual stream without requiring a real visual stream.

3 Audio-Visual Network to Super-Resolve Speech

The overview of our proposed audio-visual speech SR network is depicted in Figure 2. Given the LR speech signal, $S_{lr} = \{lr_1, lr_2, ..., lr_M\}$ and the corresponding visual frames, $V_{real} = \{V_1, V_2, ..., V_K\}$, our goal is to generate the HR speech signal, $S_{hr} = \{hr_1, hr_2, ..., hr_N\}$. During inference, if the frame sequence V_{real} is not available, we synthetically generate the frames $V_{synth} = \{v_1, v_2, ..., v_K\}$ from the LR speech input S_{lr}, as we will shortly discuss.

3.1 The Architecture

As discussed in Section 1, when the speech resolution is low, the loss of information is so paramount, that the semantic details of speech are almost completely lost. In such cases, we show that the visual stream can aid in recovering the content, thereby improving the quality, and coherence of super-resolved speech. Our proposed audio-visual model comprises three modules: (i) Speech Encoder, (ii) Visual Encoder, and (iii) Speech Decoder. We elaborate on each of these modules below.

Speech Encoder We consider a 1-second segment of LR speech S_{lr} and perform linear interpolation to upsample the LR speech to the target resolution (S_{lu}). We do this upsampling step to use the same architecture irrespective of the input resolution. We apply short-time Fourier transform (STFT) to convert the raw waveforms to linear spectrograms. A window length of 25ms with a hop length of 10ms sampled at 16kHz is considered for computing STFT. The obtained complex STFT of dimension $(T, 257)$ is decomposed into the magnitude and the phase components and normalized in the range of $[0, 1]$. We concatenate these components along the frequency axis to create $(T, 514)$ dimension representation that acts as input to the speech encoder. The speech encoder is a series of residual 1D convolution layers which processes these time-frequency representations and generates speech embeddings $(T, 600)$.

Visual Encoder We extract the visual features from the visual stream input V_{real} using the visual encoder. We design our visual encoder to process the input frames of dimension $(\frac{T}{4}, 3, 96, 96)$ by gradually reducing the spatial dimension to $(\frac{T}{4}, 600, 1, 1)$ using a stack of 3D convolution layers with residual connections. Our visual encoder is similar to the visual stream of "Perfect Match" model [9]. It captures the short-range motion information using a temporal receptive field of 5 frames in the first convolution layer. The output of the visual encoder is upsampled 4-times along the temporal axis using nearest neighbour interpolation to match the spectrogram temporal dimension T. Thus, we finally obtain the visual embeddings of dimension $(T, 600)$. Note that although the visual encoder takes real-visual stream, V_{real} as input during training, it can accept either V_{real} or the synthetic visual stream, V_{synth} during inference. We provide a detailed discussion to generate V_{synth} in Section 3.2 below.

Speech Decoder The speech decoder aims to generate a residual mask, which is added to the input spectrogram S_{lu} to obtain the output spectrogram. Initially, we fuse the learned speech

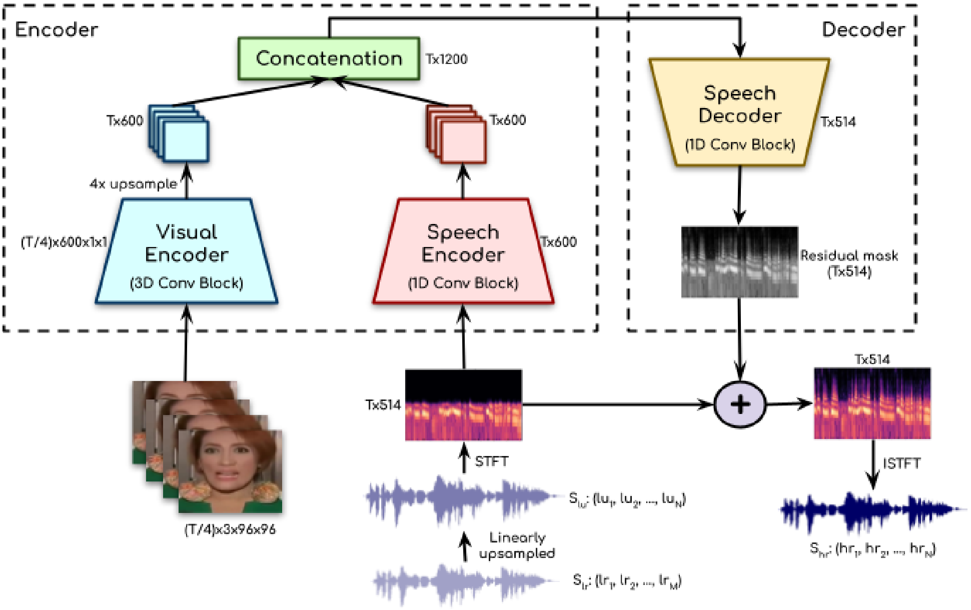

Figure 2: We propose an audio-visual network for solving speech super-resolution at large scale-factors (8× and 16×). Our SR model comprises three major components: (i) visual encoder, (ii) speech encoder, and (iii) speech decoder. The visual encoder ingests a sequence of frames, processes them and generates the visual embeddings. The speech encoder takes the spectrogram representation from the linearly upsampled speech signal to create the speech embeddings. These learned visual and speech embeddings are then fused and subsequently processed by the speech decoder. Our network outputs a residual mask which is added to the input spectrogram to generate realistic, high-quality (16kHz) speech signals.

and visual embeddings in the latent space to form $(T, 1200)$-dimension features. The decoder, which consists of 1D convolution layers, ingests these features and outputs a residual mask of dimension $(T, 514)$. Our experiments use the addition mask to get the spectrogram, as we found that the output quality is significantly better than using multiplicative masks (see Table 6 for comparison). The mean absolute error ($L1$) between the generated, and the ground-truth HR spectrograms is used as the loss to train our network. Finally, we use the inverse-STFT (ISTFT) to obtain the speech from the generated spectrograms.

3.2 Speech Super-Resolution using Pseudo-Visual Stream

3.2.1 Synthetic Generation of Frames

Our audio-visual model requires frontal or near-frontal talking-face videos of the speaker to be available as input. However, we observe that there can be situations, especially in real-world applications, where the visual stream is corrupted, unreliable, or not present altogether. The videos where the lip movements are occluded, out-of-focus or even out-of-sync with the speech cannot be considered as the visual stream input. In such cases, we propose to synthetically generate the visual stream using our pseudo-visual model. Note that our speech SR network is trained only using the real visual stream (as it is available during training), but can ingest the pseudo-visual stream during testing. This clearly demonstrates our SR network's capability and robustness, which adapts well to synthetic data during inference.

3.2.2 Student-Teacher Setup

To synthesize the visual stream from the low-resolution speech input, we adapt a student-teacher setup inspired by [14]. Specifically, the student model takes the LR speech as input and generates the frames with an objective to match the teacher model's output, as shown in Figure 3. The pre-trained teacher model Wav2Lip [28] produces accurate lip movements from the HR speech and an identity image. The student model aims to mimic these precise predictions from the teacher model using the LR speech as input. As done in [14], we consider using a static identity image (here, Taylor Swift) so that the only visual changes will be in the lip and jaw region. This will enable the student model to learn the strong correspondence between speech and lip movements. We use the input representations and the architecture as done in [14] to train the student model.

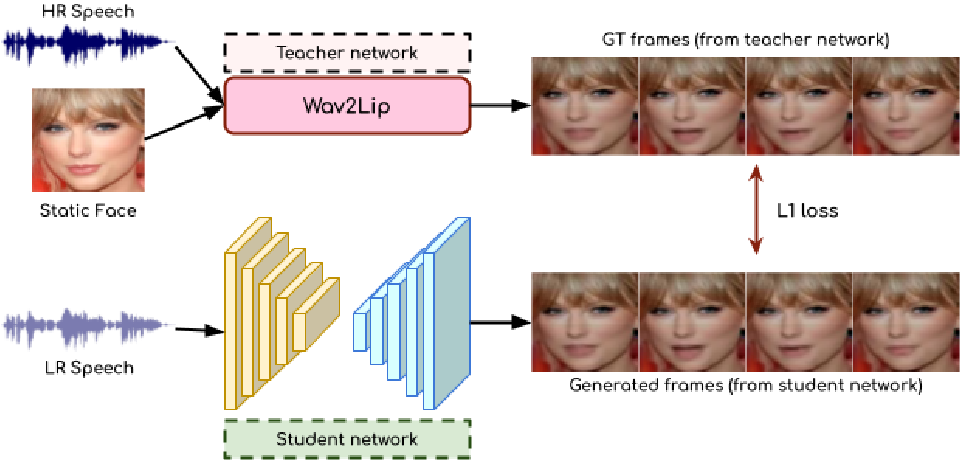

Figure 3: We demonstrate the applicability of our proposed SR network by synthesizing the lip movements in cases where the visual stream is not present. We set up a student-teacher network to generate the visual stream from the LR speech input synthetically. The student model aims to imitate the outputs from the pre-trained teacher model (Wav2Lip [28]), which ingests the HR speech and a static identity to produce accurate lip movements.

4 Experiments and Results

4.1 Dataset and Training Settings

Dataset: We use the publicly available VoxCeleb2 [8] dataset which consists of over 1 million utterances for $\sim$ 6000 identities. This dataset is highly challenging and popular due to the wide variations in the identities, languages, and extensive vocabulary. For testing, we use the official test split from the VoxCeleb2 dataset. Note that there are no overlaps between the identities used in the training and the testing set; thus, evaluating on it demonstrates the generalisation ability of our model on completely unseen identities.
Training Setup: We train our network by randomly sampling a 1-second segment of audio at 16kHz and its corresponding video frames from the VoxCeleb2 train set. The linear spectrograms extracted from the audio waveform correspond to $T = 100$ timesteps for a 1-second

segment. The corresponding frames are considered at 25 FPS and are resized to 96×96 before feeding to the visual encoder. We perform all our experiments at scale-factors of 4×, 8× and 16× with the fixed output resolution of 16kHz. The network is trained using the Adam Optimizer [16] with a learning rate of 10^{-3} and batch size of 32 and stop the training when the validation loss plateaus. In our experiments, the model was trained for 50 epochs.

4.2 Results

We now present the results of our audio-visual speech super-resolution for scale-factors of 4×, 8× and 16×. We start by discussing the various existing approaches that we use for comparison. This is then followed by the quantitative evaluation (Section 4.2.1) along with details of the different metrics used. Finally, we also conduct human evaluation (Section 4.2.2) to highlight the real-world applicability of our approach.

Comparison: We start our comparisons with standard "linearly interpolated" outputs. Next, since the existing works in the speech SR literature are limited to lower scale-factors, we train them on the same dataset as our model at all the scale-factors for a fair comparison. Additionally, we also train an audio-only (AO) baseline of our network by discarding the visual stream input. Thus, we compare against the following models: (i) Linear interpolation, (ii) DNN [22], (iii) U-Net [18], (iv) TFiLM [5], (v) NU-Wav [20], and (vi) AO baseline.

4.2.1 Quantitative Evaluation

Evaluation Metrics: We use several popular speech metrics for measuring the quality of our speech generations. We report Perceptual Evaluation of Speech Quality (PESQ) [29] which estimates the perceptual quality of the generated speech. To evaluate the intelligibility of speech, we compute Short-Time Objective Intelligibility (STOI) [30] and Extended Short-Time Objective Intelligibility (ESTOI) [15]. Finally, as done in most of the speech SR works [5, 18, 20], we also report the Log-spectral Distance (LSD) [13] metric.

Table 1: Quantitative comparison of different approaches at scale-factors of 4×, 8× and 16×. Our method outperforms the existing audio-only approaches by a large margin, illustrating the benefits from the visual stream.

Scale factor	**Method**	Linear	DNN [22]	U-Net [18]	TFiLM [5]	NU-Wav [20]	AO baseline	**Ours (pseudo)**	**Ours**
4×	PESQ↑	3.289	3.304	3.318	3.342	3.397	3.363	3.416	**3.429**
	STOI↑	0.871	0.888	0.904	0.889	0.892	0.912	0.916	**0.917**
	ESTOI↑	0.739	0.819	0.825	0.837	0.855	0.843	0.861	**0.869**
	LSD↓	6.112	6.012	6.004	5.803	5.801	5.799	**5.686**	5.694
8×	PESQ↑	2.330	2.243	2.268	2.275	2.219	2.399	2.401	**2.814**
	STOI↑	0.756	0.749	0.765	0.771	0.774	0.804	0.818	**0.832**
	ESTOI↑	0.590	0.638	0.667	0.681	0.663	0.705	0.721	**0.755**
	LSD↓	10.79	7.681	7.325	6.830	9.541	6.220	6.014	**5.069**
16×	PESQ↑	1.842	1.639	1.651	1.654	1.526	1.925	2.188	**2.237**
	STOI↑	0.550	0.653	0.671	0.684	0.598	0.702	0.726	**0.762**
	ESTOI↑	0.327	0.432	0.480	0.551	0.482	0.593	0.614	**0.651**
	LSD↓	11.405	9.306	8.993	8.082	9.780	7.841	6.601	**5.500**

Table 1 shows the results at scale-factors of 4×, 8× and 16×. We can observe that at

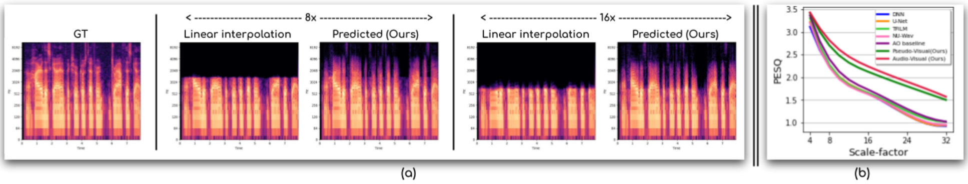

Figure 4: (a) Spectrograms of the ground-truth (GT), linearly upsampled speech, and our predicted speech. We can see that our network can reconstruct the LR speech, which is close to the GT speech even at large scale-factors. (b) Performance comparison (metric: PESQ) at different scale-factors. At higher scale-factors, the gap in the performance of "audio-only" and "audio-visual" methods emphasizes the importance of the visual stream at larger scales.

smaller scale-factors like 4×, the performance of all the approaches are very similar; the boost obtained using the visual stream is not significant. However, as the scale-factor increases, all the audio-only methods struggle to recover plausible speech outputs. At higher scale-factors of 8× and 16×, our method outperforms the other methods by a large margin, especially in perceptual quality. It is interesting to note that our pseudo-visual model not only surpasses all the current techniques, but is also very close to our approach that uses the real-visual stream. This validates the precise lip shape generations of our pseudo-visual network. Sample spectrograms shown in Figure 4 (a) depicts that our model successfully reconstructs the high-frequency elements even from very low-resolution inputs.

How does the performance vary when the scale-factor increases? In Figure 4 (b), we compare the performance of different models at various scale-factors: 4×, 8×, 16×, 24× and 32×. We can clearly notice the gap between the existing audio-only approaches and our proposed model. This difference in performance increases with the increase in scale-factor. Although there is room for improvement at scale-factors of 24× and 32×, the impact and the usefulness of the visual stream at these scales is impressive. It allows the model to recover the lost information at larger scale-factors that are otherwise much harder by solely using the audio modality.

Computation comparison: In Table 2, we compare the number of parameters and the inference time for all the models. Except for the NU-Wav [20], the parameters of our "audio-visual" model is similar to other "audio-only" models. It is to be noted that although NU-Wav has fewer parameters compared to ours, in terms of performance, we surpass NU-Wav by a large margin, especially at higher scale-factors. For comparing the inference time, we process a 1-second audio segment on a single NVIDIA Geforce RTX 2080Ti GPU. As we can see from the table, our audio-visual model is faster (2^{nd} best) compared to most of the existing audio-only models. This is mainly because all the other approaches (except AO baseline) operate at the waveform level, whereas we take the spectrogram approach which is considerably faster and also better in terms of performance.

Table 2: Comparison of the model size (in million parameters) and the inference time (in seconds). Our "audio-visual" model has similar parameters compared to most of the "audio-only" approaches, with a very low inference time.

	DNN [22]	U-Net [18]	TFiLM [5]	NU-Wav [20]	AO baseline	**Ours (pseudo)**	**Ours**
# params (M)↓	69.9	70.9	68.2	3.0	8.1	90.0	69.3
inf. time (sec)↓	1.113	1.268	0.971	2.921	0.638	0.929	0.873

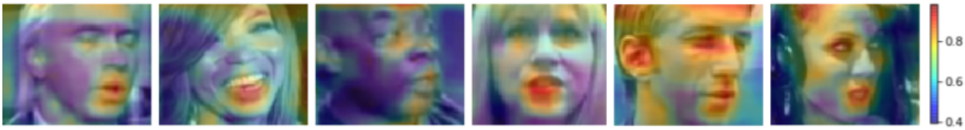

Figure 5: Activation maps of the visual encoder for different identities. Although our model is highly attentive to the lip region, the contributions from other facial areas such as eyes and cheeks are also noteworthy.

4.2.2 Human Evaluation

To assess the perceptual quality of our speech generations, we conduct a human study. We randomly select 15 samples from the test set of VoxCeleb2 dataset [8] and generate the super-resolved signals at scale-factor of 16×. The outputs from our approach and all the comparison methods are played in random order. We ask 30 participants to rate each of these speech samples on a scale of 1-5 based on: (a) Quality and (b) Intelligibility. The participant group consists of people in the age groups of 20-50 and has a nearly equal male-female ratio. The mean opinion scores (MOS) are reported in Table 3. Inline with our quantitative evaluations, our method generates speech which is largely preferred over the other methods.

Table 3: Mean opinion scores of different methods based on: (i) Quality and (ii) Intelligibility. Our method generates plausible speech outputs with higher perceptual satisfaction.

Measure	Linear	TFiLM [5]	NU-Wav [20]	AO baseline	**Ours (pseudo)**	**Ours**
Quality	2.057	2.571	2.343	2.643	3.152	**3.415**
Intelligibility	1.928	2.685	2.369	2.599	3.064	**3.282**

5 Ablation Studies

We perform several ablation experiments to analyze various aspects of our model. All the experiments are conducted for 16× SR on the VoxCeleb2 test set [8].

5.1 What kind of Visual Input is the Best?

We analyze different forms of the visual input: (i) the lower half of the face containing lip and jaw region and (ii) full face. Providing the full face performs better, as observed in Table 4. This is also reflected by the activation map in Figure 5 which shows that the facial regions like the eyes, cheeks, and forehead also play a crucial role along with the significant attention on the lip and jaw regions.

Table 4: Feeding full face to the visual encoder achieves better performance.

Method	PESQ↑	STOI↑	ESTOI↑	LSD↓
Lower half	2.425	0.743	0.638	5.633
Full face	2.237	0.762	0.651	5.500

5.2 Robustness to Noise

We show the robustness of our network in handling the noisy inputs. We add the Gaussian noise at three SNR levels of 5dB, 10dB, and 15dB to the LR speech input. As we can observe from Table 5, our model can generate plausible speech outputs even for severely degraded speech inputs.

Table 5: Our model is robust to noisy inputs and generates plausible speech outputs.

Noise level	PESQ↑	STOI↑	ESTOI↑	LSD↓
5dB	2.035	0.702	0.586	6.253
10dB	2.062	0.711	0.602	6.190
15dB	2.075	0.714	0.619	6.033

5.3 Additive Mask v/s Multiplicative Mask

We investigate the performance of different types of masks used in the prior works: (i) addition mask (used in our work), (ii) multiplication mask (used in [1]), and (iii) complex ratio mask (cRM) (used in [12]). The results are reported in Table 6. We observe that although several works benefit from the popular cRM, in our case, a simple addition mask performs better than the other kinds of masks.

Table 6: Addition mask achieves better performance compared to multiplication masks.

Masks	PESQ↑	STOI↑	ESTOI↑	LSD↓
Addition (Ours)	2.237	0.762	0.651	5.500
Multiplication	2.171	0.702	0.601	6.042
cRM	2.217	0.698	0.612	5.848

5.4 Importance of the Student Network

We assess the need for a student model to synthetically generate the visual stream during inference (if the real visual stream is absent or unreliable). Table 7 shows the comparison of directly using the teacher Wav2Lip model [28], Wav2Lip trained on the LR inputs and our proposed student network. We fine-tune the teacher Wav2Lip on the VoxCeleb2 [8] dataset for fair comparison and give the linearly upsampled speech signal as input (Wav2Lip takes speech inputs at 16kHz). As we can observe in Table 7, directly using the teacher model fails to generate plausible speech; this is evident as this network was not intended to work on LR inputs. The teacher model trained (from scratch) on LR inputs also gives poor performance. The best results are obtained using our proposed student model, which learns to imitate the accurate teacher model's output, thus validating our claim of student-teacher setup.

Table 7: Our student network yields the best performance compared to other alternatives.

Pseudo-visual models	PESQ↑	STOI↑	ESTOI↑	LSD↓
Teacher Wav2Lip [28]	1.012	0.637	0.553	8.628
Wav2Lip trained on LR	1.684	0.690	0.581	7.647
Student network (Ours)	2.237	0.762	0.651	5.500

6 Conclusion

In this work, we present the first audio-visual network for super-resolving speech signals. While the previous works were restricted to 4× SR limiting their practical applicability, our method effectively super-resolves at higher factors of 8× and 16×. We emphasize the importance of the visual stream in handling very low-resolution inputs and remarkably improving the generated speech quality. We also show the real-world applicability of our method in handling "in-the-wild" speech signals without an associated visual stream. Our designed pseudo-visual model accurately synthesizes the lip movements solely from the LR speech input. Our method achieves a considerable boost over the state-of-the-art audio-only approaches in quantitative metrics and user studies. We believe our work takes a significant step forward in the audio-visual space.

References

[1] T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. In *INTERSPEECH*, 2018.

[2] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[3] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, 2020.

[4] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, and et.al Battenberg. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 173–182. JMLR.org, 2016.

[5] 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.

[6] Y. Cheng, D. O'Shaughnessy, and P. Mermelstein. Statistical recovery of wideband speech from narrowband speech. *IEEE Trans. Speech Audio Process.*, 2:544–548, 1994.

[7] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *British Machine Vision Conference*, 2017.

[8] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[9] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3965–3969, 2019.

[10] Shaojin Ding, Tianlong Chen, Xinyu Gong, Weiwei Zha, and Zhangyang Wang. Autospeech: Neural architecture search for speaker recognition, 2020.

[11] P. Ekstrand. Bandwidth extension of audio signals by spectral band replication. In *Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA '02*, 2002.

[12] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37, 2018. doi: 10.1145/3197517.3201357.

[13] A. Gray and J. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24:380–391, 1976.

[14] Sindhu B. Hegde, K.R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. Visual speech enhancement without a real visual stream. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1926–1935, January 2021.

[15] Jesper Jensen and Cees Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24:1–1, 11 2016. doi: 10.1109/TASLP.2016.2585878.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Keisuke Kinoshita, Tsubasa Ochiai, Marc Delcroix, and Tomohiro Nakatani. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7009–7013. IEEE, 2020.

[18] Volodymyr Kuleshov, S. Enam, and S. Ermon. Audio super resolution using neural networks. *ICLR Workshops*, abs/1708.00853, 2017.

[19] Erik Larsen and R. Aarts. Audio bandwidth extension: Application of psychoacoustics, signal processing and loudspeaker design. 2004.

[20] Junhyeok Lee and Seungu Han. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *INTERSPEECH*, 2021.

[21] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777, 2014. doi: 10.1109/TASLP.2014.2304637.

[22] Kehuang Li, Z. Huang, Yong Xu, and Chin-Hui Lee. Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech. In *INTERSPEECH*, 2015.

[23] Alastair H Moore, P Peso Parada, and Patrick A Naylor. Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures. *Computer Speech & Language*, 46:574–584, 2017.

[24] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[25] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[26] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. Segan: Speech enhancement generative adversarial network. pages 3642–3646, 08 2017. doi: 10.21437/Interspeech.2017-1428.

[27] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 484–492. Association for Computing Machinery, 2020. doi: 10.1145/3394171.3413532.

[29] Antony Rix, John Beerends, Michael Hollier, and Andries Hekstra. Perceptual evaluation of speech quality (pesq): A new method for speech quality assessment of telephone networks and codecs. volume 2, pages 749–752 vol.2, 2001. doi: 10.1109/ICASSP.2001.941023.

[30] Cees Taal, Richard Hendriks, R. Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. pages 4214 – 4217, 04 2010. doi: 10.1109/ICASSP.2010.5495701.

[31] Ruilin Xu, Rundi Wu, Yuko Ishiwaka, Carl Vondrick, and Changxi Zheng. Listening to sounds of silence for speech denoising. In *Advances in Neural Information Processing Systems*, volume 33, pages 9633–9648. Curran Associates, Inc., 2020.