Supplemental Materials for 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

We provide supplemental material to the original paper [8].

A. Technical Details of Facestar Dataset

Video was captured using two synchronized OV2312 1600 x 1300 RGBIr cameras running at 60 fps. A custom camera aggregator sends video streams over USB to PC where they are recorded. Audio was captured using a custom 3D printed microphone array with 7 DPA 4060 prepolarized condenser microphones. The microphone signals are recorded by an 8 channel RME OctaMic XTC analog to digital converter running at 48kHz. The 8th channel records a shared IRIG timecode signal from a Meinberg syncbox, which is used to synchronize the video and audio subsystems. Note that in this work we use a single camera and a single microphone as inputs of our model.

Both participants signed a consent form for data usage and publication, which have undertaken internal legal and ethical review.

B. Extended Results Tables

Tables 4 and 5 show the full evaluation of models using objective metrics on the Facestar and YouTube-Lip2Wav datasets respectively. In addition to the objective metrics described in the main text, we have also provided results for speech-to-reverberation modulation energy ratio (**SRMR**) and composite measures of speech quality (**CSIG, CBAK, COVL**). Our approach consistently outperforms the baselines.

C. Extended Ablation Results

Table 6 shows extended ablation results of our model under different signal-to-noise ratios (SNR). Here, the noise added to the clip is sampled from Audioset [4], and we do not include interfering speakers or reverberation. We find that the vision modality is useful for enhancing speech, even without an interfering speaker. In particular, there is an increasing performance gap between audio-visual and audio-only models at SNRs below 40dB.

	Decoded from GT Mel-Spectrograms	Ground Truth
PESQ ↑	2.49504	5.0
STOI ↑	0.88038	1.0
MCD ↓	1.68988	0.0
Mel-Spec-Dist ↓	0.00069	0.0
CSIG ↑	4.26524	5.0
CBAK ↑	2.53051	5.0
COVL ↑	3.37861	5.0

Table 1. **Objective metrics evaluated on clean synthesized speech.** For PESQ, STOI, CSIG, CBAK, and COVL, higher is better. For MCD and Mel-Spec-Dist, lower is better. See text for details.

GT recordings	Synthesis from GT mel-spectrograms	Can not tell
21.3%	10.7%	68.0%

Table 2. **Perceptual Evaluation**. Participants were presented two video clips and asked to tell which of the two sounds more natural.

D. Importance of Human Evaluation Studies

Although the objective metrics shown in the main paper and Tables 4 and 5 are widely used in literature, it is important to note that no objective metric precisely reflects how humans perceive speech quality [5]. In particular, approaches to speech enhancement based on speech synthesis, such as ours, may generate many different waveform signals that are perceptually similar to the ground truth. Objective metrics that are relative (*i.e.*, that compare denoised audio waveform to ground truth audio waveform) may not accurately reflect the performance of these approaches.

To demonstrate, we use HiFi-GAN [6] to synthesize clean speech from ground truth mel-spectrograms. We evaluate the synthesized speech using objective metrics (Table 1) as well as human evaluation studies (Table 2). While the objective metrics would suggest that the synthesized speech is degraded compared to the ground truth speech, as shown in Table 1, human participants could not differentiate between the synthesized speech and the ground truth

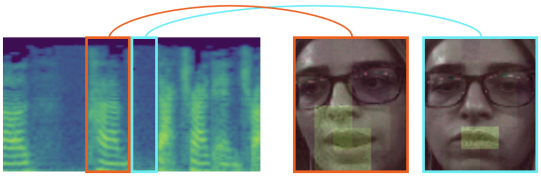

Figure 1. **Illustration of Visual Attention.** Figure shows mel-spectrogram of denoised speech, and images overlaid with heatmaps showing where the visual model attends to.

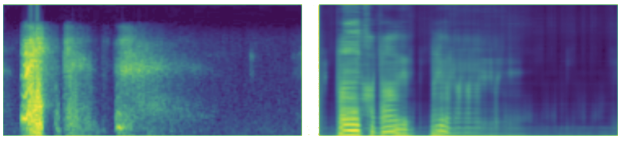

Figure 2. **Effect of Speech Codebook**. (a) Mel-spectrogram representation of noise. (b) Mel-spectrogram representation of the same noise from (a) after being converted to and reconstructed from speech codes. The reconstructed noise resembles plausible vocal sounds from the target speaker. See supplemental video for examples.

speech based on their quality, as shown in Table 2. This demonstrates the importance of human evaluation studies for evaluating speech enhancement methods, in particular, those based on speech synthesis.

E. Visual Attention of Model

The results in the main section demonstrate the importance of the visual modality for speech enhancement in our model, particularly in the presence of interfering speakers. To determine the regions of the video that the visual subnetwork attends to, we systematically zero out a 16 by 16 region of pixels of the video over 5 frames and compute the increase in error on the mel-spectrogram. We scale and overlay the results over the video to produce a heatmap of visual attention of the model. The results show that the model primarily attends to the lip motion of the speaker, as shown in Figure 1.

F. Importance of Discrete Speech Codebook

Beyond enabling the efficiency of our approach, mapping audio-visual inputs to a sequence of discrete codes from a speech codebook prevents unwanted noise from the input from being propagated through the decoder and synthesized into the output, since the codebook is trained on clean speech only and its capacity is bottle-necked. To demonstrate, we take noise audio clips N and find the sequence of codes that most closely generates this audio, *i.e.*,

Model	PESQ ↑ (Narrow Band)	STOI ↑
Speaker: Chemistry Lectures		
Lip2Wav [7]	1.300	0.416
Ours	1.956	0.807
Speaker: Chess Lectures		
Lip2Wav [7]	1.400	0.418
Ours	1.834	0.774
Speaker: Deep Learning Lectures		
Lip2Wav [7]	1.671	0.282
Ours	1.871	0.705
Speaker: Ethical Hacking Lectures		
Lip2Wav [7]	1.367	0.369
Ours	1.970	0.722
Speaker: Hardware Security Lectures		
Lip2Wav [7]	1.290	0.446
Ours	1.831	0.690

Table 3. **Comparison to Video-To-Speech Synthesis Models on YouTube-Lip2Wav Dataset [7].** Our approach outperforms Lip2Wav, the video-to-speech synthesis approach of [7]. For PESQ (narrow band) and STOI, higher is better. Results for Lip2Wav are copied directly from [7].

we optimize:

$$\min_{\mathbf{Z}} ||\mathcal{D}(\mathbf{Z}) - \textbf{melspec}(N)||_2 \quad (1)$$

Figure 2 shows an example result of this optimization: a noise clip (Figure 2(a)) is mapped to a sequence of codes that synthesizes a plausible sound from the target speaker, even though the codes are selected to reproduce the noise.

G. Comparison to Video-to-Speech Synthesis

While Figure 3 of the main text demonstrates that our synthesis approach to speech enhancement is driven by the visual modality, incorporating audio is important for ensuring faithfulness of the speaker's pitch. As shown in Table G, our approach described in Section 3 of the main paper produces significantly higher-quality results than the state-of-the-art video-to-speech synthesis model of [7]. Note that uses the narrow band version of PESQ evaluation, whereas our evaluation in the main text uses the wide band.

Model	PESQ ↑	STOI ↑	SRMR ↑	F-SNR ↑	MCD ↓	CSIG ↑	CBAK ↑	COVL ↑	Mel-ℓ_2 ↓
Speaker 1									
Demucs [1]	1.159	0.476	6.023	4.424	5.192	2.405	1.628	1.686	0.0127
AV-Masking [3]	1.233	0.567	6.744	5.510	4.990	2.472	1.710	1.752	0.00878
AV-Mapping [2]	1.293	0.596	4.850	1.223	4.831	1.012	1.038	1.001	0.00521
Ours	**1.384**	**0.651**	**9.184**	**7.409**	**3.453**	**3.079**	**1.837**	**2.177**	**0.00458**
Speaker 2									
Demucs [1]	1.34	0.632	6.128	6.781	4.815	2.707	**1.767**	1.956	0.00861
AV-Masking [3]	1.280	0.620	6.581	6.473	5.380	2.459	1.679	1.777	0.00985
AV-Mapping [2]	**1.373**	0.657	4.373	4.381	4.940	1.514	1.304	1.200	**0.00659**
Ours	1.325	**0.672**	**6.752**	**7.236**	**4.179**	**2.764**	1.705	**1.959**	0.00667

Table 4. **Quantitative Evaluation of Audio-Visual Speech Separation and Enhancement on Facestar Dataset.** Our approach consistently outperforms the baselines. For PESQ, STOI, SRMR, F-SNR, CSIG, CBAK, COVL, higher is better. For MCD and Mel-ℓ_2, lower is better.

Model	PESQ ↑	STOI ↑	SRMR ↑	F-SNR ↑	MCD ↓	CSIG ↑	CBAK ↑	COVL ↑	Mel-ℓ_2 ↓
Speaker: Chemistry Lectures									
Demucs [1]	1.308	0.731	6.986	6.992	5.343	2.544	1.652	1.858	0.01610
AV-Masking [3]	1.319	0.717	7.107	7.187	5.598	2.476	1.591	1.810	0.01160
AV-Mapping [2]	1.485	0.751	5.969	7.295	4.413	2.695	1.726	1.975	0.00568
Ours	**1.503**	**0.807**	**10.076**	**9.247**	**3.725**	**3.130**	**1.885**	**2.264**	**0.00486**
Speaker: Chess Lectures									
Demucs [1]	**1.526**	**0.820**	3.158	**10.189**	**4.069**	**3.178**	**1.986**	**2.335**	0.00625
AV-Masking [3]	1.426	0.756	**4.056**	9.128	5.069	2.868	1.833	2.113	0.00703
AV-Mapping [2]	1.360	0.706	3.485	7.466	4.876	2.533	1.643	1.829	0.00521
Ours	1.393	0.774	2.904	9.546	4.309	3.006	1.867	2.162	**0.00467**
Speaker: Deep Learning Lectures									
Demucs [1]	1.336	0.566	8.957	7.049	4.756	2.636	1.578	1.895	0.01236
AV-Masking [3]	**1.539**	0.651	9.759	8.101	5.099	2.758	1.728	2.069	0.00975
AV-Mapping [2]	1.472	0.627	6.145	7.032	4.626	2.359	1.435	1.721	0.00646
Ours	1.539	**0.705**	**11.529**	**8.684**	**4.308**	**2.946**	**1.816**	**2.170**	**0.00585**
Speaker: Ethical Hacking Lectures									
Demucs [1]	1.321	0.613	9.110	7.366	4.382	2.715	1.760	1.947	0.00924
AV-Masking [3]	1.387	0.655	9.407	7.539	4.888	2.700	1.651	1.955	0.00841
AV-Mapping [2]	1.390	0.630	7.101	6.308	4.273	2.488	1.571	1.778	0.00561
Ours	**1.491**	**0.722**	**13.104**	**8.664**	**3.592**	**3.073**	**1.832**	**2.211**	**0.00475**
Speaker: Hardware Security Lectures									
Demucs [1]	1.424	0.631	11.481	6.621	5.069	2.627	1.704	1.937	0.01062
AV-Masking [3]	**1.521**	0.66431	11.045	7.412	5.180	2.699	1.701	2.022	0.00950
AV-Mapping [2]	1.379	0.589	5.090	6.357	5.027	2.430	1.508	1.765	0.00800
Ours	1.487	**0.690**	**12.875**	**7.862**	**4.428**	**2.878**	**1.704**	**2.105**	**0.00747**

Table 5. **Quantitative Evaluation of Audio-Visual Speech Separation and Enhancement on YouTube-Lip2Wav [7] Dataset.** Our approach consistently outperforms the baselines. For PESQ, STOI, SRMR, F-SNR, CSIG, CBAK, COVL, higher is better. For MCD and Mel-ℓ_2, lower is better.

SNR	0dB	10dB	20dB	30dB	40dB
Audio-Only	0.0100	0.0066	0.0048	0.0040	0.0038
No Auto-Regressive Module	0.0055	0.0041	0.0036	0.0035	0.0034
Full Model	**0.0047**	**0.0038**	**0.0034**	**0.0031**	**0.0029**

Table 6. **Extended 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.

References

[1] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, 2020. 3

[2] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *arXiv preprint arXiv:1711.08789*, 2017. 3

[3] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. *arXiv preprint arXiv:2101.03149*, 2021. 3

[4] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 1

[5] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël AP Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, et al. A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016(1):1–19, 2016. 1

[6] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *arXiv preprint arXiv:2010.05646*, 2020. 1

[7] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805, 2020. 2, 3

[8] Karren Yang, Dejan Markovic, Steven Krenn, Vasu Agrawal, and Alexander Richard. Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1