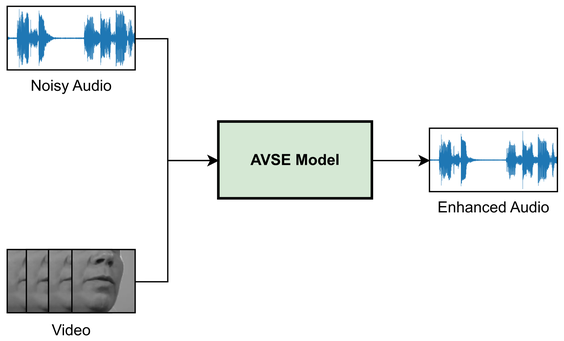
LA-VocE: Low-SNR Audio-visual Speech Enhancement using Neural Vocoders

Rodrigo Mira[1] Buye Xu[2] Jacob Donley[2] Anurag Kumar[2] Stavros Petridis[1,3] Vamsi Krishna Ithapu[2] Maja Pantic[1,3]

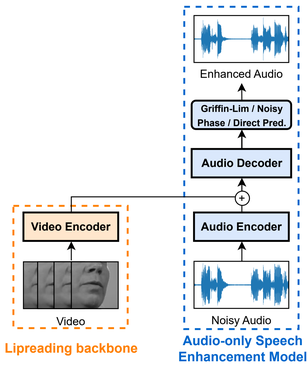
[1]iBUG, Imperial College London, UK [2]Meta Reality Labs Research, USA [3]Meta, UK

Motivation



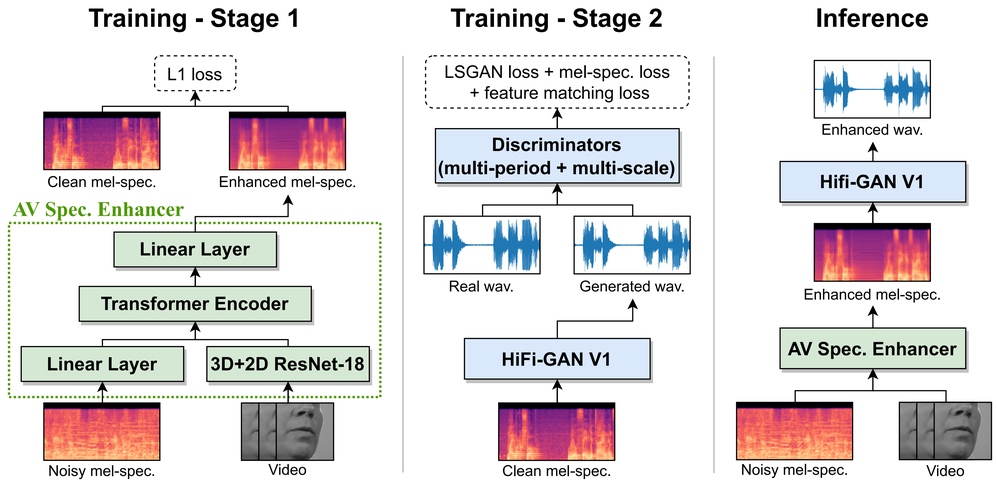
- **Audio-visual speech enhancement (AVSE)** aims to enhance audio by leveraging the speaker's lip movements.
- Can be trained on **raw unlabelled audio-visual data** by combining clean speech with noise on the fly.
- Has many applications, including **video conferencing** and **hearing augmentation** in noisy/crowded environments.

Previous Approaches



- Typically combine lipreading backbones with **existing audio-only speech enhancement models**.
- Often rely on **Griffin-Lim** or re-use the **noisy phase**.

Our Method - LA-VocE



- We propose a **two-stage approach**:
  - We train a **transformer-based spectrogram enhancer** inspired by recent audio-visual speech recognition models.
  - Then, we train a **neural vocoder (HiFi-GAN)** on the same corpus to generate raw audio from spectrograms.
  - Finally, we **combine** both during inference to perform end-to-end enhancement.
- We train our model by combining clean speech from **AVSpeech (∼4,700 hours, 11+ languages)** with noise from the **DNS Challenge noise dataset (∼70,000 samples, ∼150 classes)**.

Noise and Interference Study

		PESQ-WB i ↑					ESTOI i ↑				
	SNR (dB)	5	0	-5	-10	-15	5	0	-5	-10	-15
SIR (dB)	5	0.970	0.876	0.715	0.486	0.245	0.269	0.316	0.356	0.375	0.362
	0	0.904	0.795	0.630	0.411	0.210	0.327	0.354	0.375	0.378	0.355
	-5	0.789	0.679	0.511	0.319	0.136	0.386	0.394	0.397	0.383	0.349
	-10	0.617	0.523	0.405	0.248	0.092	0.429	0.426	0.414	0.388	0.344
	-15	0.438	0.383	0.289	0.195	0.081	0.443	0.433	0.414	0.381	0.330

		PESQ-WB i ↑					ESTOI i ↑				
	# noises	1	2	3	4	5	1	2	3	4	5
# spk.	1	0.709	0.642	0.601	0.580	0.557	0.396	0.402	0.404	0.404	0.403
	2	0.602	0.553	0.511	0.497	0.482	0.396	0.398	0.397	0.395	0.393
	3	0.539	0.490	0.462	0.455	0.431	0.390	0.390	0.388	0.387	0.384

- We study our model's performance by **varying the SNR/SIR of the evaluation samples (shown on the left)**, and **varying the number of background noises and interfering speakers (shown on the right)**.
- Remarkably, LA-VocE yields **substantial improvements in intelligibility (ESTOI i) even in the most extreme scenarios**, such as -15 dB SNR/SIR (on the left) or 5 background noises + 3 interfering speakers (on the right).

Comparison with Other Works

Method	Input	MCD i ↓	PESQ-WB i ↑	ViSQOL i ↑	STOI i ↑	ESTOI i ↑
Noise condition 1 (1 backg. noise at 0 dB SNR + 1 interf. speaker at 0 dB SIR)						
GCRN	A	0.410	0.044	0.093	-0.052	-0.038
AV-GCRN	AV	-1.193	0.394	0.499	0.220	0.235
AV-Demucs	AV	-5.581	0.738	0.688	0.270	0.298
MuSE	AV	-5.528	0.787	0.679	0.276	0.299
VisualVoice	AV	-3.781	0.606	0.645	0.249	0.270
LA-VocE (audio-only)	A	-3.189	0.248	0.135	0.055	0.047
LA-VocE	AV	**-6.653**	**0.931**	**1.100**	**0.294**	**0.333**
Noise condition 2 (3 backg. noises at -5 dB SNR + 2 interf. speakers at -5 dB SIR)						
GCRN	A	-0.416	-0.010	0.163	-0.015	-0.015
AV-GCRN	AV	-1.354	0.096	0.398	0.234	0.214
AV-Demucs	AV	-5.548	0.274	0.426	0.308	0.300
MuSE	AV	-5.314	0.297	0.409	0.308	0.289
VisualVoice	AV	-3.388	0.164	0.367	0.253	0.237
LA-VocE (audio-only)	A	-2.817	0.056	0.087	0.066	0.043
LA-VocE	AV	**-6.863**	**0.511**	**0.700**	**0.379**	**0.397**
Noise condition 3 (5 backg. noises at -10 dB SNR + 3 interf. speakers at -10 dB SIR)						
GCRN	A	-0.414	-0.015	0.210	-0.020	-0.005
AV-GCRN	AV	-1.263	-0.043	0.217	0.171	0.139
AV-Demucs	AV	-4.866	0.013	0.298	0.262	0.230
MuSE	AV	-4.185	0.011	0.242	0.231	0.182
VisualVoice	AV	-2.518	-0.045	0.248	0.181	0.160
LA-VocE (audio-only)	A	-1.982	-0.015	0.073	0.032	0.008
LA-VocE	AV	**-6.170**	**0.159**	**0.447**	**0.371**	**0.358**

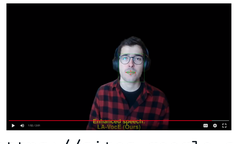
- We achieve **state-of-the-art speech enhancement performance on 3 challenging low-SNR noise conditions**.

Spec. Inversion Comparison

Method	Train. corp.	MCD i ↓	PESQ-WB i ↑	ViSQOL i ↑	STOI i ↑	ESTOI i ↑	Spec. MSE i ↓
Griffin-Lim	-	-6.805	0.333	**0.806**	0.311	0.318	-7.855
Noisy phase	-	-6.640	0.461	0.721	0.305	0.310	-7.901
HiFi-GAN	VCTK	-6.570	0.384	0.655	0.374	0.388	-7.773
HiFi-GAN	LJSpeech	-6.601	0.432	0.670	0.370	0.382	-7.825
HiFi-GAN	AVSpeech	**-6.863**	**0.511**	0.700	**0.379**	**0.397**	**-7.939**

- Our HiFi-GAN (trained on AVSpeech) **outperforms pre-trained models, Griffin-Lim and noisy phase reconstruction**.

Project Page (with Demos!)



https://sites.google.com/view/la-voce-avse