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

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]Imperial College London [2]Meta Reality Labs Research [3]Meta AI
rs2517@imperial.ac.uk

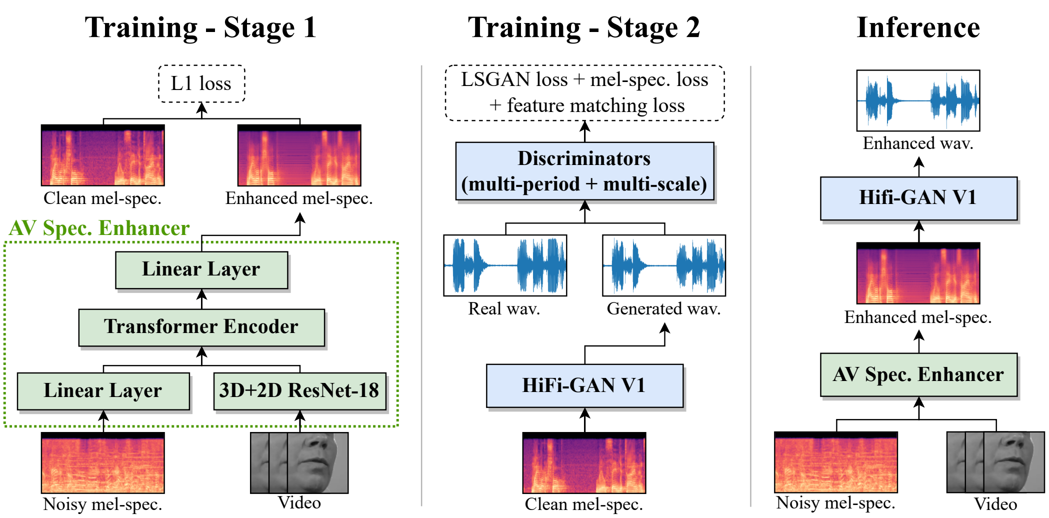

Figure 1: Summary of LA-VocE's two-stage training approach and inference procedure.

1. Introduction

Speech enhancement is a well-established signal processing task that aims to remove background noise from a speech signal. In past years, novel deep learning models have been leveraged to push the state-of-the-art in the field [24], but generally focus on high-SNR (signal-to-noise ratio) scenarios and often neglect the potential presence of overlapping speech [22, 2]. These challenges have drawn interest to the idea of exploiting the visual modality to improve performance in these more extreme scenarios - this is known as audio-visual speech enhancement (AVSE). This approach is particularly promising given the newfound ubiquity of video conferencing, as well as the recent success of video-to-speech models [14], which are able to synthesize speech using only the speaker's silent lip movements.

Contemporary AVSE models are typically designed by adapting a U-Net-based architecture from an audio-only speech enhancement approach [2, 22], adding a visual stream to model the speaker's lip movements and combining the acoustic and visual features in the model's bottleneck [4, 16, 5]. While this approach is grounded in existing research, it neglects the emergence of new transformer-based audio-visual encoders, such as the ones presented in recent audio-visual speech recognition approaches [12, 20]. Furthermore, these models often rely on masking techniques [16, 5] or re-use the phase from the noisy signal [4, 7], which works well for high-SNR scenarios, but becomes less viable when the input signal is extremely noisy and the original signal is barely perceptible.

With these two issues in mind, we propose LA-VocE (**L**ow-SNR **A**udio-visual **Voc**oder-based speech **E**nhancement), featuring a new two-stage approach to tackle this challenge. First, we train an audio-visual spectrogram enhancer which consists of a ResNet-based visual encoder and a linear acoustic encoder, followed by a large transformer that learns to predict the clean spectrogram from the combined audio-visual embedding. Then, we adapt an existing neural vocoder (HiFi-GAN [9]) to generate the waveform corresponding to each spectrogram and train it on the same corpus. Finally, during inference, we combine these two models to perform audio-visual speech enhancement from raw video and audio to raw waveform.

[†]Work done during internship at Meta

2. Methodology

We summarize our methodology in Figure 1. In stage 1, our spectrogram enhancer receives video of the cropped mouth and encodes it using a 2D ResNet-18 [6] preceded by a 3D Convolutional layer (as in [17, 15, 14]), and also the noisy log-mel spectrogram, which is encoded into acoustic features using a linear layer. Then, these two sets of features are concatenated along the channel dimension (the visual features are temporally upsampled to match the acoustic features) and fed into the transformer. We apply a transformer encoder [23] which is composed of a front-end embedding layer and 12 transformer blocks with attention dimension 768, feedforward dimension 3072, and 12 attention heads. The resulting features are projected into the predicted spectrogram using a linear layer. We train this model using an L1 loss between the predicted and clean spectrograms.

In stage 2, we adopt a state-of-the-art neural vocoder, HiFi-GAN [9], to generate raw audio from our predicted spectrograms. In particular, we use HiFi-GAN V1, which contains 12 ResBlocks that sequentially upsample the log-mel spectrogram into the final waveform. The model is trained via a multi-period discriminator (MPD), which analyzes the generated waveform across different periods, and a multi-scale discriminator (MSD), which discriminates downsampled versions of the waveforms. In this stage, our training loss is a combination of the LSGAN (Least Squares Generative Adversarial Network) loss [13], an L1 loss between the real and generated spectrograms, and a feature matching loss for the discriminators [11].

3. Experiments

Datasets, pre-processing, and augmentation. We train our model by combining clean speech with randomly sampled noise and interfering speech (clean speech that is added to the background as noise) on the fly. We draw clean speech (as well as interfering speech) from AVSpeech [3], which is known as one of the largest publicly available audio-visual speech datasets. It contains roughly 4,700 hours of video, featuring 11+ languages. To sample noise, we use the DNS Challenge noise dataset [18], which contains roughly 70,000 noise clips spanning around 150 classes (*e.g.* car noises, background music). Due to computational constraints, we sample only 1 % of the test set for AVSpeech and use this as our evaluation set.

We control the level of background noise via the signal-to-noise ratio (SNR) and the level of interfering speech via the signal-to-interference ratio (SIR):

$$\text{SNR} = \frac{P_{signal}}{P_{noise}}, \qquad \text{SIR} = \frac{P_{signal}}{P_{interference}}, \quad (1)$$

where P refers to the power of each waveform. During training, SNR and SIR vary randomly and independently between 5 and -15 dB. During evaluation, we instead design three noise conditions where the SNR and SIR are fixed at 0, -5, and -10 dB, the number of background noises is set to 1, 3, and 5, and the number of interfering speakers is set to 1, 2, and 3, for noise levels 1, 2, and 3, respectively.

Method	Input	MCD i ↓	PESQ-WB i ↑	ViSQOL i ↑	STOI i ↑	ESTOI i ↑
Noise condition 1 (1 background noise at 0 dB SNR + 1 interfering speaker at 0 dB SIR)						
GCRN [22]	A	0.410	0.044	0.093	-0.052	-0.038
AV-GCRN [22]	AV	-1.193	0.394	0.499	0.220	0.235
AV-Demucs [2]	AV	-5.581	0.738	0.688	0.270	0.298
MuSE [16]	AV	-5.528	0.787	0.679	0.276	0.299
VisualVoice [5]	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 background noises at -5 dB SNR + 2 interfering speakers at -5 dB SIR)						
GCRN [22]	A	-0.416	-0.010	0.163	-0.015	-0.015
AV-GCRN [22]	AV	-1.354	0.096	0.398	0.234	0.214
AV-Demucs [2]	AV	-5.548	0.274	0.426	0.308	0.300
MuSE [16]	AV	-5.314	0.297	0.409	0.308	0.289
VisualVoice [5]	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 background noises at -10 dB SNR + 3 interfering speakers at -10 dB SIR)						
GCRN [22]	A	-0.414	-0.015	0.210	-0.020	-0.005
AV-GCRN [22]	AV	-1.263	-0.043	0.217	0.171	0.139
AV-Demucs [2]	AV	-4.866	0.013	0.298	0.262	0.230
MuSE [16]	AV	-4.185	0.011	0.242	0.231	0.182
VisualVoice [5]	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**

Table 1: Comparison between LA-VocE and other speech enhancement methods for different noise conditions.

Evaluation metrics To evaluate the quality of our results, we apply a set of well-established speech metrics: Mean Cepstral Distance (MCD) [10], wideband PESQ (PESQ-WB) [19], Virtual Speech Quality Objective Listener (ViSQOL) [1], Short-Time Objective Intelligibility (STOI) [21], and its extension ESTOI [8]. We denote improvements between noisy and enhanced audio with 'i', *e.g.* PESQ-WB i.

Results We present our results in Table 1, after training all models under equivalent conditions on the datasets presented above. Firstly, it is clear that the audio-only methods fail to yield any noticeable improvements in any scenario. This is expected since, without visual information, these methods cannot accurately distinguish interfering speech from the target signal. Moving on to the audio-visual methods, under noise condition 1, LA-VocE achieves state-of-the-art performance across all metrics, outperforming previous methods based on spectral mapping (AV-GCRN [22]), waveform reconstruction (AV-Demucs [2]), and masking (MuSE [16] and VisualVoice [5]). When we move on to noise condition 2, it is clear that the other models, despite being trained on the same data, feature a sharp decline in performance, while LA-VocE continues to yield substantial improvements, particularly on STOI and ESTOI. Finally, on noise condition 3, other approaches are unable to yield noticeable improvements due to the exceptionally high level of noise corrupting the original signal. LA-VocE, on the other hand, is able to yield large improvements on most metrics even in this extreme scenario, demonstrating its robustness to low-SNR settings.

4. Conclusion

In conclusion, we present a new two-stage approach for audio-visual speech enhancement under low-SNR conditions entitled LA-VocE. We train and evaluate our model on AVSpeech [3] and compare our results with previous audio-only and audio-visual enhancement models using multiple objective metrics. In our results, we show that LA-VocE consistently outperforms existing methods across three different noise conditions.

Acknowledgements

Only non-Meta authors conducted any of the dataset pre-processing (no dataset pre-processing took place on Meta's servers or facilities).

References

[1] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines. ViSQOL v3: An open source production ready objective speech and audio metric. In *QoMEX*, pages 1–6. IEEE, 2020. 2

[2] A. Défossez, G. Synnaeve, and Y. Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, pages 3291–3295. ISCA, 2020. 1, 2

[3] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4):112, 2018. 2, 3

[4] A. Gabbay, A. Shamir, and S. Peleg. Visual speech enhancement. In *Interspeech*, pages 1170–1174. ISCA, 2018. 1

[5] R. Gao and K. Grauman. VisualVoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, pages 15495–15505. IEEE, 2021. 1, 2

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE, 2016. 2

[7] J. Hou, S. Wang, Y. Lai, Y. Tsao, H. Chang, and H. Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Trans. Emerg. Top. Comput. Intell.*, 2(2):117–128, 2018. 1

[8] J. Jensen and C. H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *ACM Trans. Audio Speech Lang. Process.*, 24(11):2009–2022, 2016. 2

[9] J. Kong, J. Kim, and J. Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *NeurIPS*, volume 33, pages 17022–17033, 2020. 1, 2

[10] R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Pacific Rim Conf. on Commun. Comput. and Signal Process.*, volume 1, pages 125–128 vol.1, 1993. 2

[11] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1558–1566. JMLR.org, 2016. 2

[12] P. Ma, S. Petridis, and M. Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP*, pages 7613–7617. IEEE, 2021. 1

[13] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2813–2821. IEEE, 2017. 2

[14] R. Mira, A. Haliassos, S. Petridis, B. W. Schuller, and M. Pantic. SVTS: scalable video-to-speech synthesis. In H. Ko and J. H. L. Hansen, editors, *Interspeech*, pages 1836–1840. ISCA, 2022. 1, 2

[15] R. Mira, K. Vougioukas, P. Ma, S. Petridis, B. W. Schuller, and M. Pantic. End-to-end video-to-speech synthesis using generative adversarial networks. *IEEE Transactions on Cybernetics*, pages 1–13, 2022. 2

[16] Z. Pan, R. Tao, C. Xu, and H. Li. Muse: Multi-modal target speaker extraction with visual cues. In *ICASSP*, pages 6678–6682. IEEE, 2021. 1, 2

[17] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. End-to-end audiovisual speech recognition. In *ICASSP*, pages 6548–6552. IEEE, 2018. 2

[18] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke. The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. In *Interspeech*, pages 2492–2496. ISCA, 2020. 2

[19] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, pages 749–752. IEEE, 2001. 2

[20] B. Shi, W. Hsu, K. Lakhotia, and A. Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *ICLR*. OpenReview.net, 2022. 1

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *ICASSP*, pages 4214–4217. IEEE, 2010. 2

[22] K. Tan and D. Wang. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:380–390, 2020. 1, 2

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2

[24] D. Wang and J. Chen. Supervised speech separation based on deep learning: An overview. *IEEE ACM Trans. Audio Speech Lang. Process.*, 26(10):1702–1726, 2018. 1