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

ABSTRACT

Audio-visual speech enhancement aims to extract clean speech from a noisy environment by leveraging not only the audio itself but also the target speaker's lip movements. This approach has been shown to yield improvements over audio-only speech enhancement, particularly for the removal of interfering speech. Despite recent advances in speech synthesis, most audio-visual approaches continue to use spectral mapping/masking to reproduce the clean audio, often resulting in visual backbones added to existing speech enhancement architectures. In this work, we propose LA-VocE, a new two-stage approach that predicts mel-spectrograms from noisy audio-visual speech via a transformer-based architecture, and then converts them into waveform audio using a neural vocoder (HiFi-GAN). We train and evaluate our framework on thousands of speakers and 11+ different languages, and study our model's ability to adapt to different levels of background noise and speech interference. Our experiments show that LA-VocE outperforms existing methods according to multiple metrics, particularly under very noisy scenarios.

***Index Terms*—** Audio-visual speech enhancement, speech separation, speech synthesis, neural vocoder, transformer.

1. INTRODUCTION

Speech enhancement, defined as the extraction of clean speech from a noisy signal, is a well-established signal processing task which has benefited greatly from the advent of deep learning [1]. Recently-proposed models excel at denoising and dereverberation [2, 3], but often struggle with very low signal-to-noise ratios (SNR) [4]. Furthermore, audio-only methods struggle to accurately remove background speech, as they are limited in the information they can use to distinguish it from the target signal. These limitations have drawn researchers to leverage visual cues of the target speaker's lip movements as additional supervision – an approach known as audio-visual speech enhancement (AVSE). This can be particularly valuable for applications such as video conferencing, streaming, recording and hearing augmentation in a crowded and/or noisy environment, where the target speaker's video stream can help the model enhance their speech. This method may also be leveraged to improve speech recognition in low-SNR conditions. Furthermore, the recent success of video-to-speech synthesis [5, 6], where the audio is reproduced using only silent video, highlights the importance of the visual modality and shows a promising direction for audio-visual speech enhancement in very low-SNR conditions.

Recent AVSE methods are often based on U-Nets [7–9], inspired by their audio-only counterparts [2, 3, 10], or simple convolutional networks [11], frequently combined with LSTMs [12]. Existing speech enhancement models are typically combined with a video encoder which extracts visual features and concatenates them with the acoustic features to perform audio-visual enhancement. These approaches draw from speech enhancement literature, but fail to leverage state-of-the-art audio-visual encoders [13, 14]. Most methods estimate (either directly or via a mask) the magnitude and phase of the clean spectrogram, which are converted into waveform using the inverse Short-Time Fourier Transform (iSTFT) [7, 9, 11], while others attempt to perform enhancement in the time domain directly [8]. Both of these reconstruction techniques rely on very accurate predictions, which can be difficult to achieve, especially in low-SNR environments where audio supervision is unreliable. Recent works in audio-only [15, 16] and audio-visual [12] speech enhancement have introduced neural vocoders as an alternative synthesis method, but choose to focus on high-SNR scenarios where this reconstruction technique is likely to have a lesser impact. Alternatively, new works introduce neural codecs [17] for waveform synthesis but focus heavily on achieving compressed representations, which is not a priority for most speech enhancement frameworks.

To address these shortcomings, we propose a new two-stage approach for audio-visual speech enhancement entitled **L**ow-SNR **A**udio-visual **Voc**oder-based Speech **E**nhancement (LA-VocE). First, we train an audio-visual spectrogram enhancer, which receives noisy speech and video of the cropped mouth, and aims to predict a clean spectrogram. This model features a ResNet-18-based visual encoder [18] and a large transformer encoder [19] to model the temporal correlations in the audio-visual features, and is trained using an L1 loss between the real and predicted mel-spectrogram magnitudes. We then train a neural vocoder (HiFi-GAN V1 [20]) to predict waveform audio from clean mel-spectrograms on the same corpus. This fully convolutional model is trained using a mixture of adversarial and comparative losses, with an ensemble of eight discriminators operating on multiple periods and scales. During inference, the enhancer and the vocoder are combined to perform end-to-end audio-visual speech enhancement.

Our contributions are as follows: **(1)** We present a new audio-visual speech enhancement approach that combines a transformer-based spectrogram enhancer with our version of HiFi-GAN V1. **(2)** We train our model to remove background noise and speech on the challenging AVSpeech dataset. **(3)** We compare our approach with previous state-of-the-art models, and show that it significantly outperforms all methods across all metrics and noise conditions. **(4)** We study our model's ability to generate clean audio for varying levels of noise and interference and find that it consistently achieves improvements in speech intelligibility. **(5)** We measure our trained vocoder's effectiveness against other spectrogram inversion approaches and observe that it significantly outperforms other methods.

[*]Work done during internship at Meta

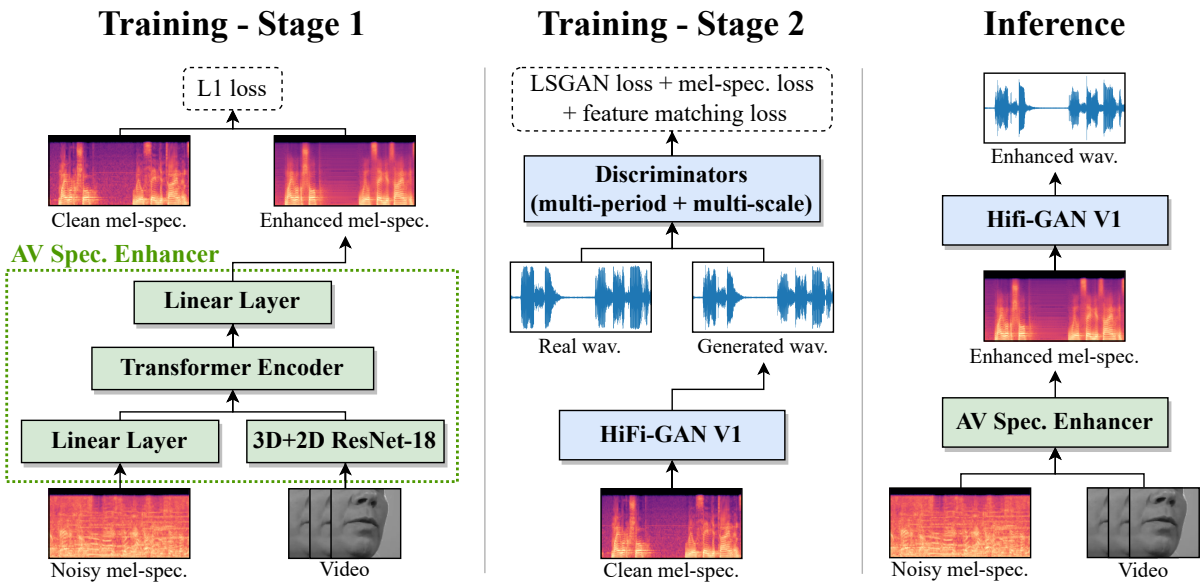

Fig. 1. Summary of LA-VocE's two-stage training approach and inference procedure.

2. METHODOLOGY

2.1. Audio-visual spectral enhancement

LA-VocE is summarized in Fig. 1. The first stage in our framework consists of training an audio-visual spectrogram enhancer. This model extracts visual features using a 2D ResNet-18 [18] with a 3D convolutional stem (as in [5, 6, 21, 22]), and acoustic features using a single linear layer. The video features are then upsampled (via nearest neighbors interpolation) to match the audio features' frame rate, and the features from the two modalities are concatenated along the channel dimension. The fused audio-visual features are fed into the transformer encoder [19] - the largest component in the network. This module comprises an initial embedding layer, with a linear layer followed by relative positional encoding [23], and 12 transformer encoder blocks, where the attention dimension, feedforward dimension, and the number of attention heads are 768, 3072, and 12, respectively. Finally, these features are decoded via a linear layer into the predicted mel-frequency spectrogram. We train the model by applying an L1 Loss:

$$\mathcal{L}_1 = \|s_{clean} - E(s_{noisy}, v)\|_1, \tag{1}$$

where s_{clean} and s_{noisy} are the clean and noisy mel-spectrograms, respectively, v is the video of the speaker's lip movements and E is our audio-visual spectrogram enhancer.

2.2. Waveform synthesis

The second stage in our method involves training a neural vocoder to convert the enhanced spectrograms into waveform audio. We use HiFi-GAN [20], which upsamples the spectrogram gradually using a set of transposed convolutions. In particular, we opt for HiFi-GAN V1, the largest HiFi-GAN variant, which features 12 ResBlocks with hidden size 512, amounting to 13.92 million parameters. As proposed in [20], HiFi-GAN is trained via a multi-period discriminator (MPD), composed of five convolutional sub-discriminators which analyze the waveform along different periods (i. e., every 2, 3, 5, 7 and 11 samples), and a multi-scale discriminator (MSD), consisting of one sub-discriminator for the raw audio and two sub-discriminators that receive downsampled versions of the same waveform (via $2\times$ and $4\times$ average pooling). Our training objective (as in the original HiFi-GAN) combines the Least Squares Generative Adversarial Network (LSGAN) loss [24] with an L1 loss on the mel-spectrogram magnitudes and a feature matching loss [25]:

$$\mathcal{L}_G = \alpha_1 \mathcal{L}_{G_{adv}} + \alpha_2 \mathcal{L}_{spec} + \alpha_3 \mathcal{L}_{FM}, \tag{2}$$

$$\mathcal{L}_{G_{adv}} = \sum_{i=1}^{N_D} (D_i(G(s_{clean})) - 1)^2, \tag{3}$$

$$\mathcal{L}_{spec} = \|m(w_{clean}) - m(G(s_{clean}))\|_1, \tag{4}$$

$$\mathcal{L}_{FM} = \sum_{i=1}^{N_D} \sum_{l=1}^{N_{L_i}} \frac{\|D_i^l(w_{clean}) - D_i^l(G(s_{clean}))\|_1}{d_i^l}, \tag{5}$$

$$\mathcal{L}_D = \sum_{i=1}^{N_D} (D_i(w_{clean}) - 1)^2 + D_i(G(s_{clean}))^2, \tag{6}$$

where $\mathcal{L}_G$ is the generator loss, $\mathcal{L}_D$ is the discriminator loss, $\mathcal{L}_{G_{adv}}$ is the generator's adversarial loss, $\mathcal{L}_{spec}$ is the mel-spectrogram loss, $\mathcal{L}_{FM}$ is the feature matching loss, G is the generator (HiFi-GAN V1), D_i is the i-th discriminator, N_D is the number of discriminators, w_{clean} is the clean waveform, m is the function that computes the mel-spectrogram, N_{L_i} is the number of layers in discriminator i, and D_i^l and d_i^l refer to the features extracted from layer l / discriminator i and their dimension, respectively. Loss coefficients α_1, α_2 and α_3 are set to 1, 45, and 2, respectively, as in [20]. After both stages of training, the spectrogram enhancer and neural vocoder are combined during inference, as shown in Fig. 1.

3. EXPERIMENTAL SETUP

3.1. Datasets, pre-processing, and augmentation

Our experiments focus on AVSpeech [26], one of the largest publicly available audio-visual speech datasets. It contains ~4,700 hours of video, featuring ~150,000 different subjects and 11+ languages. The scale and heterogeneity of the data make for a substantially more challenging than many commonly-used corpora such as GRID [7, 12, 27] or Facestar [12], which are recorded in studios.

We sample background noise from the Deep Noise Supression challenge [28] noise dataset. It contains roughly 70,000 noise clips, amounting to around 150 classes, ranging from music to machine sounds. Both datasets are split into training, validation and testing sets using a 80 – 10 – 10 % ratio. Due to the computational cost of computing the evaluation metrics, we randomly sample 1 % of the AVSpeech testing set, amounting to 1552 samples, and use this as the evaluation set for our experiments. We add two types of corruption to the clean speech: background noise (denoted 'noise') and background speech (denoted 'interference'). The corruption level is controlled by the Signal-to-Noise Ratio (SNR) and the Signal-to-Interference Ratio (SIR):

$$\text{SNR} = \frac{P_{signal}}{P_{noise}}, \qquad \text{SIR} = \frac{P_{signal}}{P_{interference}}, \tag{7}$$

where P refers to the power of each waveform. The interfering speech is also obtained from AVSpeech. During training, the SNR and SIR are independently randomly sampled between 5 and -15 dB. The number of background noises and interfering speakers in each sample varies randomly from 1 to 5 and 1 to 3, respectively. During validation, we propose three different noise conditions to compare with other methods, ranging from least to most noisy. Noise conditions 1 (low), 2 (medium), and 3 (high) feature 1, 3, and 5 background noises at 0, -5, and -10 dB SNR, and 1, 2, and 3 background speakers at 0, -5, and -10 dB SIR, respectively.

The noisy and clean signals are normalized via peak normalization, and are converted into log-scale mel-spectrograms using the following parameters: frequency bin size and Hann window size 1024, hop size 256 and 80 mel bands. The audio sampling rate is 16 kHz and the video frame rate is 25 frames per second (fps). To model the speaker's lip movements, we extract the 96×96 grayscale mouth Region Of Interest (ROI) from each video, following [6, 22]. To augment our data, we apply random cropping, random horizontal flipping, random erasing, and time-masking, as in [6].

3.2. Evaluation metrics

We evaluate our results using a set of five objective speech metrics. To measure speech quality, we use Mean Cepstral Distance (MCD) [29], the wideband version of Perceptual Evaluation of Speech Quality (PESQ-WB) [30], and Virtual Speech Quality Objective Listener (ViSQOL) [31]. To measure intelligibility we use Short-Time Objective Intelligibility (STOI) [32] and its extended version ESTOI [33]. Finally, in our spectrogram inversion comparison, we also measure the mean squared error between the STFT magnitudes of each signal and refer to this as Spec. MSE. Following other works [8, 10], we denote improvements between noisy and enhanced speech metrics with the lowercase 'i', e. g., PESQ-WB i.

3.3. Comparison models

We compare our results with two recent AVSE models: VisualVoice [9], a complex spectral masking approach originally proposed for speech separation that we adapt to perform enhancement, and Multi-modal Speaker Extraction (MuSE) [8], a feature masking approach based on Conv-TasNet [10]. To provide a broader comparison with other reconstruction techniques, we also adapt two recent speech enhancement models for AVSE - Gated Convolutional Recurrent Network (GCRN) [2] and Demucs [3]. We achieve this by adding a visual stream (3D front-end + ResNet-18, as in our model) which encodes the video into temporal features that are concatenated with the audio features from the original audio encoder (preceding the LSTM/GLSTM). We refer to these models as AV-GCRN and AV-Demucs. We also compare with the original audio-only GCRN and an audio-only version of LA-VocE to highlight the importance of the visual stream. All models are implemented based on official open-source code.

3.4. Training details

We train our spectrogram enhancer for 150 epochs using AdamW [34] with learning rate 7×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.98$ and weight decay 3×10^{-2}. We increase the learning rate for the first 15 epochs using linear warmup, and then apply a cosine decay schedule [35]. To train MuSE, we replace the original SI-SDR objective [8] with the loss from Demucs (L1 + multi-resolution STFT [3]), as we find this increases training stability and yields better results. We train an audio-only version of LA-VocE by removing the visual encoder and changing the attention dimension and the number of heads in the transformer to 256 and 8, respectively. We train HiFi-GAN for roughly 1 million iterations on AVSpeech using AdamW with learning rate 2×10^{-4}, $\beta_1 = 0.8$, $\beta_2 = 0.99$ and weight decay 1×10^{-2}, decaying the learning rate by a factor of 0.999 every epoch.

4. RESULTS

4.1. Comparison with other works

We compare with previous state-of-the-art methods in Table 1, and present a demo of these results on our project website[1]. For noise condition 1, LA-VocE outperforms other approaches in quality and intelligibility, achieving significant improvements across all metrics. Indeed, even in this less noisy scenario, our vocoder-based approach is able to reproduce the target speech more accurately than mask-based methods such as MuSE [8] and VisualVoice [9], which are designed for separation with one to two background speakers. Previous AVSE methods yield decreased improvements for noise condition 2, particularly for speech quality metrics such as PESQ and ViSQOL, while LA-VocE yields significant gains in quality and especially intelligibility, as indicated by ESTOI i. This shows that despite identical training conditions, previous methods adapt poorly to lower SNR / SIR conditions compared to our new model.

Finally, on the noisiest scenario (noise condition 3), it is clear that other audio-visual methods, including mapping-based approaches (AV-GCRN [2] and AV-Demucs [3]), are unable to increase speech quality, achieving effectively no improvement on PESQ-WB and small increases on other metrics. LA-VocE, on the other hand, can still achieve significant gains in all metrics, indicating that it is substantially more robust to extremely low-SNR scenarios. Notably, both audio-only models (GCRN [2] and LA-VocE) yield poor results in all scenarios - without visual information, these models cannot accurately distinguish target speech from background speech.

4.2. Noise and interference study

We study our model's performance in Table 2 by varying the SNR and SIR between 5 dB and -15 dB (as in training), while keeping the number of background noises and interfering speakers fixed at 3 and 2, respectively. On the left, we can see that PESQ-WB i peaks for higher SNR / SIR conditions and deteriorates as the noise and interference increase. This suggests that the model excels at improving speech quality for higher SNR / SIR, even with the higher PESQ baseline, but struggles to achieve substantial gains when the environment becomes too noisy. On the other hand, ESTOI i is substantially more consistent across all conditions, and is in fact higher

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

Table 1. Comparison between LA-VocE and other speech enhancement methods for different noise conditions. In the second column, "A" and "AV" stand for audio-only and audio-visual, 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 [2]	A	0.410	0.044	0.093	-0.052	-0.038
AV-GCRN [2]	AV	-1.193	0.394	0.499	0.220	0.235
AV-Demucs [3]	AV	-5.581	0.738	0.688	0.270	0.298
MuSE [8]	AV	-5.528	0.787	0.679	0.276	0.299
VisualVoice [9]	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 [2]	A	-0.416	-0.010	0.163	-0.015	-0.015
AV-GCRN [2]	AV	-1.354	0.096	0.398	0.234	0.214
AV-Demucs [3]	AV	-5.548	0.274	0.426	0.308	0.300
MuSE [8]	AV	-5.314	0.297	0.409	0.308	0.289
VisualVoice [9]	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 [2]	A	-0.414	-0.015	0.210	-0.020	-0.005
AV-GCRN [2]	AV	-1.263	-0.043	0.217	0.171	0.139
AV-Demucs [3]	AV	-4.866	0.013	0.298	0.262	0.230
MuSE [8]	AV	-4.185	0.011	0.242	0.231	0.182
VisualVoice [9]	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 2. LA-VocE's performance for different SNR / SIR conditions with 3 background noises and 2 interfering speakers.

		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

for -15 dB SNR / SIR than it is for 5 dB SNR / SIR. Indeed, LA-VocE achieves impressive improvements in intelligibility even for -15 dB SNR / SIR, where the target speech is entirely imperceptible for human listeners. This is consistent with our perceptual evaluation - LA-VocE consistently produces intelligible audio despite the noticeable artifacts for lower SNR / SIRs.

Remarkably, LA-VocE performs better for lower SIRs compared to lower SNRs, e. g., 5 dB SNR / -15 dB SIR substantially outperforms -15 dB SNR / 5 dB SIR on both metrics. This disparity is likely due to the nature of these two signals. Speech typically has a consistent frequency range, and often contains gaps that the model can easily exploit, while noise is substantially more heterogeneous, ranging from impulses to continuous noises, presenting a greater denoising challenge. We also evaluate our model's ability to perform enhancement under multiple noise sources and background speakers in Table 3, keeping the SNR / SIR at -5 dB. Unsurprisingly, we find that the best PESQ-WB i is achieved with 1 noise and 1 speaker, and becomes worse as they are increased. While it is expected that increasing the number of sources will increase the complexity of the background noise, therefore making the enhancement task more difficult, we hypothesize that the sharper drop in performance when increasing the number of speakers is related to the temporal and spectral gaps in the interference. A single stream of speech will contain pauses that will ease denoising, but these disappear as we increase the number of speakers, resembling continuous noise.

4.3. Spectrogram inversion comparison

Finally, we compare our trained HiFi-GAN with other spectrogram inversion methods in Table 4. We observe that our HiFi-GAN achieves better performance than existing pre-trained models[2] (trained on VCTK [37] and LJSpeech [38], as presented in [20]) on all six metrics, highlighting the importance of training our own vocoder on AVSpeech, rather than applying a publicly available pre-trained model as in [6]. We also compare with Griffin-Lim [36], a commonly-used spectrogram inversion algorithm, and experiment by applying iSTFT using the phase from the noisy input to reconstruct the waveform, as proposed in [7, 11]. In our experiments, both methods consistently produce artifacts that make the resulting waveforms sound noticeably more robotic than those produced by neural vocoders (this is particularly noticeable for Griffin-Lim). We show that these inversion methods yield significant drops in PESQ-WB i, STOI i, and ESTOI i, but surprisingly achieve competitive MCD i and Spec. MSE i performance, and substantially better ViSQOL i. This inconsistency likely implies that these three metrics are less sensitive to the specific artifacts introduced by these phase estimation strategies, and emphasizes the need for multiple evaluation metrics when evaluating synthesized speech.

Table 3. LA-VocE's performance for different numbers of background noises and interfering speakers (-5 dB SNR / SIR).

		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

Table 4. Comparison between different spectrogram inversion methods for LA-VocE (noise condition 2). In the upper row, "Train. corp." stands for training corpus.

Method	Train. corp.	MCD i ↓	PESQ-WB i ↑	ViSQOL i ↑	STOI i ↑	ESTOI i ↑	Spec. MSE i ↓
Griffin-Lim [36]	-	-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 [20]	VCTK	-6.570	0.384	0.655	0.374	0.388	-7.773
HiFi-GAN [20]	LJSpeech	-6.601	0.432	0.670	0.370	0.382	-7.825
HiFi-GAN [20]	AVSpeech	**-6.863**	**0.511**	0.700	**0.379**	**0.397**	**-7.939**

5. CONCLUSION

In this paper, we propose LA-VocE, a new framework for audio-visual speech enhancement under low-SNR conditions. Our method consists of two stages of training: audio-visual spectral enhancement via a transformer-based encoder, and waveform synthesis via HiFi-GAN. We train our model on thousands of hours of multilingual audio-visual speech, and find that it significantly outperforms previous state-of-the-art AVSE approaches, particularly for higher noise conditions. We study LA-VocE's performance under varying levels of noise and interference, showing that even in the noisiest scenarios our vocoder-based approach can achieve large improvements in speech intelligibility. Finally, we compare our vocoder with existing spectrogram inversion methods, highlighting the importance of training our own HiFi-GAN. In the future, we believe it would be promising to adapt our architecture for real-time synthesis, which would enable speech enhancement in live video streams.

6. ACKNOWLEDGEMENTS

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

[2] https://github.com/jik876/hifi-gan

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] 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.*, vol. 28, pp. 380–390, 2020.

[3] A. Défossez, G. Synnaeve, *et al.*, "Real time speech enhancement in the waveform domain," in *Interspeech*, ISCA, 2020, pp. 3291–3295.

[4] X. Hao, X. Su, *et al.*, "UNetGAN: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition," in *Interspeech*, ISCA, 2019, pp. 1786–1790.

[5] R. Mira, K. Vougioukas, *et al.*, "End-to-end video-to-speech synthesis using generative adversarial networks," *IEEE Transactions on Cybernetics*, pp. 1–13, 2022.

[6] R. Mira, A. Haliassos, *et al.*, "SVTS: scalable video-to-speech synthesis," in *Interspeech*, ISCA, 2022, pp. 1836–1840.

[7] A. Gabbay, A. Shamir, *et al.*, "Visual speech enhancement," in *Interspeech*, ISCA, 2018, pp. 1170–1174.

[8] Z. Pan, R. Tao, *et al.*, "Muse: Multi-modal target speaker extraction with visual cues," in *ICASSP*, IEEE, 2021, pp. 6678–6682.

[9] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in *CVPR*, IEEE, 2021, pp. 15 495–15 505.

[10] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[11] J. Hou, S. Wang, *et al.*, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 2, no. 2, pp. 117–128, 2018.

[12] K. Yang, D. Markovic, *et al.*, "Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis," in *CVPR*, IEEE, 2022, pp. 8217–8227.

[13] P. Ma, S. Petridis, *et al.*, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*, IEEE, 2021, pp. 7613–7617.

[14] B. Shi, W. Hsu, *et al.*, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *ICLR*, OpenReview.net, 2022.

[15] Z. Du, X. Zhang, *et al.*, "A joint framework of denoising autoencoder and generative vocoder for monaural speech enhancement," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1493–1505, 2020.

[16] H. Li and J. Yamagishi, "Noise tokens: Learning neural noise templates for environment-aware speech enhancement," in *Interspeech*, ISCA, 2020, pp. 2452–2456.

[17] N. Zeghidour, A. Luebs, *et al.*, "Soundstream: An end-to-end neural audio codec," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 495–507, 2022.

[18] K. He, X. Zhang, *et al.*, "Deep residual learning for image recognition," in *CVPR*, IEEE, 2016, pp. 770–778.

[19] A. Vaswani, N. Shazeer, *et al.*, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[20] J. Kong, J. Kim, *et al.*, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.

[21] S. Petridis, T. Stafylakis, *et al.*, "End-to-end audiovisual speech recognition," in *ICASSP*, IEEE, 2018, pp. 6548–6552.

[22] P. Ma, S. Petridis, *et al.*, "Visual speech recognition for multiple languages in the wild," *CoRR*, vol. abs/2202.13084, 2022.

[23] Z. Dai, Z. Yang, *et al.*, "Transformer-xl: Attentive language models beyond a fixed-length context," in *ACL*, Association for Computational Linguistics, 2019, pp. 2978–2988.

[24] X. Mao, Q. Li, *et al.*, "Least squares generative adversarial networks," in *ICCV*, IEEE, 2017, pp. 2813–2821.

[25] A. B. L. Larsen, S. K. Sønderby, *et al.*, "Autoencoding beyond pixels using a learned similarity metric," in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 48, JMLR.org, 2016, pp. 1558–1566.

[26] A. Ephrat, I. Mosseri, *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, p. 112, 2018.

[27] M. Cooke, J. Barker, *et al.*, "An audio-visual corpus for speech perception and automatic speech recognition (l)," *The Journal of the Acoustical Society of America*, vol. 120, pp. 2421–4, 2006.

[28] C. K. A. Reddy, V. Gopal, *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Interspeech*, ISCA, 2020, pp. 2492–2496.

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

[30] A. W. Rix, J. G. Beerends, *et al.*, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, IEEE, 2001, pp. 749–752.

[31] M. Chinen, F. S. C. Lim, *et al.*, "ViSQOL v3: An open source production ready objective speech and audio metric," in *QoMEX*, IEEE, 2020, pp. 1–6.

[32] C. H. Taal, R. C. Hendriks, *et al.*, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP*, IEEE, 2010, pp. 4214–4217.

[33] 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.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, OpenReview.net, 2019.

[35] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *ICLR*, OpenReview.net, 2017.

[36] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. on Acoust., Speech, and Sig. Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[37] J. Yamagishi, C. Veaux, *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," in *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2012.

[38] K. Ito and L. Johnson, *The LJ speech dataset*, `https://keithito.com/LJ-Speech-Dataset/`, 2017.