Interspeech 2024
1-5 September 2024, Kos, Greece

RT-LA-VocE: Real-Time Low-SNR Audio-Visual Speech Enhancement

Honglie Chen[*1], *Rodrigo Mira*[*2], *Stavros Petridis*[1,2], *Maja Pantic*[1,2]

[1]Meta AI, UK [2]Imperial College London, UK

hongliechen@meta.com, {rs2517, stavros.petridis04, m.pantic}@imperial.ac.uk

Abstract

In this paper, we aim to generate clean speech frame by frame from a live video stream and a noisy audio stream without relying on future inputs. To this end, we propose RT-LA-VocE, which completely re-designs every component of LA-VocE, a state-of-the-art non-causal audio-visual speech enhancement model, to perform causal real-time inference with a 40 ms input frame. We do so by devising new visual and audio encoders that rely solely on past frames, replacing the Transformer encoder with the Emformer, and designing a new causal neural vocoder *C-HiFi-GAN*. On the popular AVSpeech dataset, we show that our algorithm achieves state-of-the-art results in all real-time scenarios. More importantly, each component is carefully tuned to minimize the algorithm latency to the theoretical minimum (40 ms) while maintaining a low end-to-end processing latency of 28.15 ms per frame, enabling real-time frame-by-frame enhancement with minimal delay.

Index Terms: Audio-visual, speech enhancement, real-time, emformer, neural vocoder

1. Introduction

When conversing in the real world, our speech is often mixed with other undesirable signals, such as noise from cars on the road, or another conversation happening nearby. For this reason, the concept of noise reduction in real time has long been considered an attractive prospect for modern speech transmission frameworks [1]. With this in mind, several deep learning-based methods have recently been proposed to tackle this challenge by adapting existing non-causal models [2] or designing new architectures tailored specifically for real-time enhancement [3, 4]. These models achieve impressive noise reduction performance in high-SNR (signal-to-noise ratio) environments but do not experiment with substantially noisier scenarios, where the performance of audio-only models often sharply deteriorates [5–7]. Furthermore, these single-channel audio-only models are unable to remove interfering speech in a multi-talker environment, limiting their potential applications.

In recent years, the increased bandwidth of modern devices has paved the way for video streaming as a ubiquitous form of communication, particularly in professional environments [8]. Compared to audio-only pipelines, video streaming adds a visual stream of the speaker's face, which can contain valuable verbal information, as demonstrated by various studies in audio-visual learning [9, 10]. In light of this, emerging approaches have borrowed from speech enhancement and lipreading literature to develop real-time audio-visual speech enhancement (AVSE) frameworks that effectively outperform their audio-only counterparts [11–14]. Most works focus on real-time inference on edge devices equipped with low-end CPUs, such as [15–17], which present a real-time AVSE model trained on studio-recorded datasets (GRID [18] and TCD-TIMIT [19]), and [20], which extends an existing real-time speech enhancement model [4] by adding a lightweight video encoder and a multi-stage fusion module. Alternatively, [21] proposes a transformed-based model running on a high-end GPU, leveraging the target speaker's facial landmarks. These real-time audio-visual approaches succeed in outperforming their audio-only counterparts, but fail to experiment with low-SNR scenarios featuring multiple interfering speakers and do not directly attempt to compete with non-causal AVSE models [11–14, 22]. More importantly, existing AVSE models are limited in terms of deployment as they benchmark inference speed on long sequences, *e.g.*, 10s in [21] and 3s in [20], and do not report the model's processing latency per frame. In fact, frame-by-frame enhancement, which is necessary for online enhancement in real-world scenarios, is almost entirely neglected by existing real-time AVSE works in both their methodology and experimental procedure.

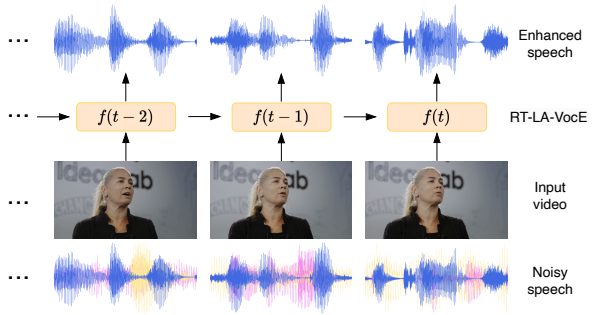

Figure 1: *RT-LA-VocE's real-time audio-visual speech enhancement approach. The model processes 40 ms frames in real time.*

To bridge the aforementioned gaps, we aim to explore real-time low-SNR audio-visual speech enhancement with minimal frame-by-frame latency (shown in Figure 1), enabling live noise reduction in real-word video streaming systems. To this end, we develop a causal AVSE model based on the state-of-the-art non-causal model LA-VocE [22], which consists of a large Transformer-based spectrogram enhancer and a neural vocoder (HiFi-GAN V1 [23]). We do so by building new, fully causal video and audio encoders, replacing the transformer with the recently proposed Emformer [24] and proposing a new vocoder – *C-HiFi-GAN* – which can synthesize raw waveforms from spectrograms without depending on future information.

Concretely, we make the following contributions: (i) We introduce a novel causal AVSE architecture designed for low-SNR conditions - RT(Real-Time)-LA-VocE; (ii) we outperform all other causal approaches and perform competitively with

*Equal contribution.

10.21437/Interspeech.2024-516

the state-of-the-art non-causal models; more importantly (iii), we reduce the algorithm latency to the minimum for real-time AVSE models, *i.e.*, 1 video frame (40 ms), so a negligible delay is introduced during real-time inference; and finally (iv), we achieve a total processing latency of 28.15 ms per frame, considerably less than the time needed to obtain the next frame (40 ms) on a server-side GPU, demonstrating the model's capability for real-world streaming applications.

2. Methodology

2.1. LA-VocE

We start by describing the multi-stage approach introduced in LA-VocE which consists of an audio-visual spectrogram enhancer and an adapted version of HiFi-GAN V1 [23]. In the first stage, given a video clip and a spectrogram, the visual and auditory representations are computed using a 2D ResNet-18 preceded by a 3D front-end layer (as in [9]) and a linear encoder, respectively. To jointly model both modalities, the two sets of features are concatenated channel-wise and passed through a 12-block Transformer encoder [25]. The output from the Transformer is then fed to a linear projection layer to generate the enhanced spectrogram. This model is trained by minimising the L1 loss between the enhanced and clean spectrograms. In the second stage, a neural vocoder is leveraged to translate the enhanced spectrogram into a waveform - HiFi-GAN V1 [23]. In short, HiFi-GAN's generator is a fully convolutional neural network consisting of 4 blocks of transposed convolutional layers and multi-receptive field fusion modules (MRF). It is trained using a combination of comparative and adversarial losses, which leverage an essemble of multi-period and multi-scale discriminators.

Limitations. The goal of this paper is to enable real-time speech enhancement with minimal latency, *i.e.*, to produce a causal model that can sequentially generate enhanced audio frame by frame from a live audio-visual stream without relying on future inputs. To achieve this, the most straightforward solution is to naïvely send frame-by-frame inputs to the original LA-VocE. However, this model is designed to rely heavily on future information along several modules, namely: the 3D convolution in the visual encoder, which leverages 2 future video frames; the STFT (short-time Fourier transform) used to compute the mel-spectrogram, whose window extends 15 ms into the future; the multi-head attention layers in the spectrogram enhancer, where the output for time-step t depends on all past and future time-steps; and the HiFi-GAN, whose convolutions gradually bleed future information into their outputs. Therefore, this naïve approach leads to severly deteriorated performance, as we will demonstrate in Table 3.

2.2. Real-time LA-VocE

To resolve these challenges, the model should be constrained to depend exclusively on past information. In this section, we describe our key changes to each module of the original LA-VocE and show our proposed RT-LA-VocE is completely causal with minimal algorithm latency. In Figure 2, we illustrate our end-to-end real-time inference pipeline in detail.

Video encoder. As proposed in [26], a convolutional layer can be made causal by adding p_{conv} padding frames to the left of the input and removing an equivalent amount of frames from the right of the output (to maintain the same output dimension):

$$p_{conv} = \lfloor \frac{k}{2} \rfloor \cdot d, \tag{1}$$

where k and d denote the kernel size and dilation of the convolution, respectively. As shown in Figure 2, we conduct this change on the 3D convolution in the video encoder, which creates a causal model that depends only on past video frames. This change reduces the algorithm latency of the video encoder from 120 ms (3 frames) to 40 ms (1 frame).

Audio encoder. LA-VocE encodes audio by converting the raw waveform into a mel-spectrogram, and applying a linear layer on the resulting magnitudes. The STFT operation used to obtain the mel-spectrogram introduces extra algorithm latency since the input raw waveform is padded on both sides by p_{stft}:

$$p_{stft} = \frac{w - h}{2}, \tag{2}$$

where w specifcies the window size (40 ms) and h specifies the hop size (10 ms), meaning that the STFT window depends on 15 ms of future information. There are two potential solutions to remove this extra latency : i) pad the left side of the waveform by $2p$ (analogous to the convolution trick mentioned above); or ii) apply a causal encoder on the raw waveform instead. We adapt the 1D ResNet-18 from [10] using the causal padding trick on all convolutions to form an entirely causal raw audio encoder. We also increase the temporal resolution from 25 Hz (one audio feature per video frame) to 100 Hz (four audio features per video frame) by adjusting the kernel size and stride in the final pooling layer. We compare these causal encoder variants in Table 2.

Emformer. In addition to the convolutional layers mentioned above, the Transformer encoder also explicitly takes into account future information via the multi-head attention operation. In order to achieve a causal temporal encoder, we replace the Transformer encoder with the recently proposed Emformer [24]. It adapts the original Transformer architecture for causal inference by computing attention over the current frame and a sequence of previous frames known as the left context, removing the attention layers' reliance on future time-steps.

C-HiFi-GAN. Similar to the 3D layer in the video encoder, the convolutions and transposed convolutions in HiFi-GAN also leak future information into the temporal representations, making the model non-causal. We design *C-HiFi-GAN* (Causal HiFi-GAN) by employing the causal padding trick for all convolutions in each MRF module of the original HiFi-GAN V1, as well as the four transposed convolutions by padding with:

$$p_{conv_{trans}} = \lfloor \frac{s}{2} \rfloor + s \bmod 2, \tag{3}$$

where s denotes the stride of the transposed convolution.

3. Experimental Setup

3.1. Datasets and evaluation metrics

Following the original LA-VocE [22], we sample clean and interfering speech from AVSpeech [27] and noise from the DNS challenge noise dataset [28]. For consistency, we use the same train/test split as the original LA-VocE. The level of speech interference (*i.e.*, audio from other speakers in the background) and background noise are controlled by the signal-to-interference ratio (SIR) and signal-to-noise ratio (SNR), respectively:

$$\text{SIR} = \frac{P_{signal}}{P_{interference}}, \qquad \text{SNR} = \frac{P_{signal}}{P_{noise}}, \tag{4}$$

where P_x refers to the power of the raw signal x. We follow the approach in [22] to crop the 96×96-sized mouth region

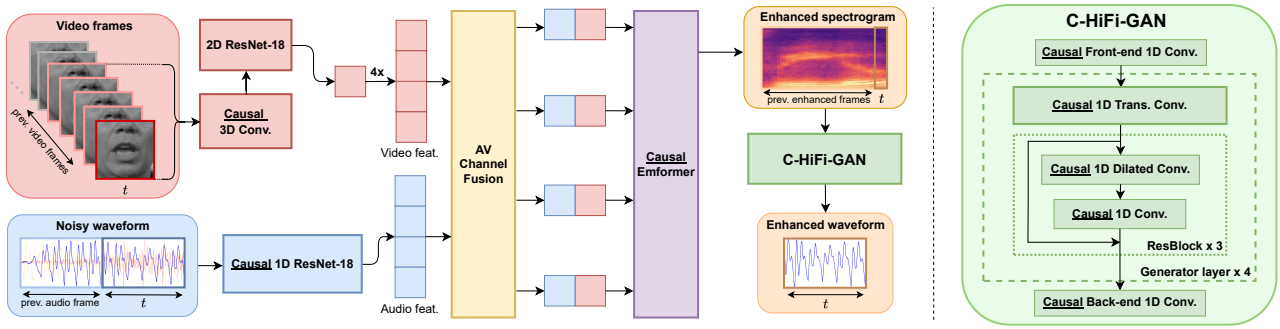

Figure 2: *Detailed overview of RT-LA-VocE's inference pipeline for each time-step t (40 ms). RT-LA-VocE receives five video frames, which are passed through our ResNet-based visual encoder, and two raw audio frames, which are encoded via our causal 1D ResNet-18. The resulting features are concatenated channel-wise and fed into the Emformer, which models the temporal dynamics with previous time-steps. This is followed by a linear layer that predicts the four enhanced spectrogram frames. Finally, these frames are combined with past predictions and fed into* C-HiFi-GAN*, which generates the corresponding waveform.*

from each video[1]. Additionally, a low-latency causal mouth cropping pipeline is created using MediaPipe [29]. We compare these two methods for real-time inference in Table 4. We measure speech quality using MCD [30], PESQ-WB [31], and ViSQOL [32], and speech intelligibility using STOI [33] and its extended version ESTOI [34]. Unlike [22], we present the raw metrics computed on the enhanced samples rather than the improvement over the noisy baseline to clearly illustrate their similarity with the clean speech in absolute terms. We evaluate all models under the same three noise conditions 1, 2, and 3 proposed in [22], featuring 1, 3, and 5 background noises at 0, -5, and -10 dB, and 1, 2, and 3 interfering speakers at 0, -5, and -10 dB SIR, respectively.

3.2. Comparison models

We compare our work with three non-causal AVSE models: LA-VocE [22], MuSE [11], and VisualVoice [12]. In addition, we adapt two speech enhancement models – GCRN [35] and Demucs [36] – for real-time audio-visual enhancement by adding a causal ResNet-based encoder to each, as in our model, and concatenating the resulting visual features with the audio features in the bottleneck, as in [22]. We use the causal version of Demucs presented in [2]. We also compare with two real-time audio-only models: GCRN and an audio-only version of RT-LA-VocE, where we remove the visual stream entirely. All models are trained on the dataset presented in Section 3.1 with optimization parameters based on the ones used for our spectrogram enhancer (described in Section 3.3). As in [22], MuSE is trained using the loss ensemble proposed in [36].

3.3. Implementation details

Architectural Details. We use window size 640, hop size 160, frequency bin size 640, and 80 mel bands to compute the mel-spectrograms. Our Emformer consists of 12 blocks featuring 12 attention heads, a hidden unit dimension of 768, a segment length of 4, a left context length of 64 and a memory bank length of 0. Note, in our experiments, no noticeable improvement was observed with memory bank lengths > 0. We use 4 blocks of transposed convolutional layers in *C-HiFi-GAN* with upsampling factors of $8\times$, $5\times$, $2\times$, and $2\times$. MRF is made of 3 residual blocks featuring dilated and non-dilated convolutional layers with kernel sizes 3, 7, and 11. During testing, each input audio frame's duration is 40 ms, *i.e.*, a single video frame.

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

Training curriculum. To train the spectrogram enhancer, we use AdamW with learning rate 7×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and weight decay 3×10^{-2}. We train for 150 epochs using linear learning rate warmup for the first 10 % of training, and applying a cosine decay schedule for the remaining epochs. Separately, we train *C-HiFi-GAN* using AdamW with learning rate 2×10^{-4}, $\beta_1 = 0.8$, and $\beta_2 = 0.99$. We train for a total of 1M steps while decaying the learning rate by a factor of 0.999 after each epoch. During training, we apply standard data augmentations such as random cropping, horizontal flipping, erasing, and time masking, as in [22]. The latencies in the following section are measured on an NVIDIA RTX 2080 Ti GPU with an Intel Core i7-9700K CPU, and averaged over 1000 inference steps.

4. Results

4.1. Comparison with the state-of-the-art

We compare our results with recent causal and non-causal speech enhancement models across three challenging noise conditions in Table 1. Overall, it is clear that RT-LA-VocE yields substantial improvements compared to the noisy baseline, halving the MCD and achieving relative improvements greater than 50 % for STOI and ESTOI throughout all 3 noise conditions. Conversely, RT-AV-GCRN and RT-AV-Demucs are only able to achieve considerable improvements for noise condition 1 and feature a sharp decline in noise conditions 2 and 3. While RT-LA-VocE does not exactly match LA-VocE's performance, it is remarkably competitive with this state-of-the-art model on both quality and intelligibility without leveraging any future information. Furthermore, RT-LA-VocE outperforms two recent non-causal models (AV-GCRN and VisualVoice) across all noise conditions, and is competitive with AV-Demucs and MuSE, achieving better performance on noise conditions 2 and 3. Finally, we highlight that both audio-only speech enhancement models fail to achieve noticeable improvements in quality or intelligibility. Indeed, as mentioned in Section 1, with only a single stream of audio these models are unable to distinguish the target signal from the interfering speech.

4.2. Ablation study

Architecture. To validate our architectural design, we first compare spectrogram variants and inversion methods in the top half of Table 2. We experiment with two different types of spectrogram as input for our linear encoder: mel-spectrogram (as in [22]) and linear spectrogram. We find that the mel-spectrogram input yields better results across all metrics. In addition, we compare *C-HiFi-GAN* with two common spectrogram inversion

Model	Input	Online	Proc. latency (ms)	Noise condition 1					Noise condition 2					Noise condition 3				
				MCD↓	PESQ↑	ViSQOL↑	STOI↑	ESTOI↑	MCD↓	PESQ↑	ViSQOL↑	STOI↑	ESTOI↑	MCD↓	PESQ↑	ViSQOL↑	STOI↑	ESTOI↑
Baseline	-	-	-	10.81	1.088	1.168	0.545	0.367	12.05	1.105	1.053	0.386	0.195	12.47	1.160	1.035	0.281	0.093
AV-GCRN	AV	✗	-	9.297	1.514	1.710	0.773	0.614	10.457	1.215	1.483	0.630	0.424	10.989	1.112	1.271	0.462	0.245
AV-Demucs	AV	✗	-	5.260	1.818	1.852	0.813	0.661	6.522	1.383	1.484	0.692	0.495	7.608	1.172	1.334	0.543	0.323
MuSE	AV	✗	-	5.282	1.875	1.847	0.821	0.666	6.736	1.402	1.462	0.694	0.484	8.285	1.171	1.277	0.512	0.275
VisualVoice	AV	✗	-	6.869	1.689	1.815	0.794	0.637	8.515	1.272	1.419	0.637	0.431	9.806	1.114	1.283	0.460	0.252
LA-VocE	AV	✗	-	4.177	2.017	2.265	0.839	0.700	5.217	1.621	1.743	0.765	0.592	6.320	1.317	1.480	0.651	0.451
RT-GCRN	A	✓	4.12±0.10	11.275	1.129	1.260	0.495	0.327	11.682	1.102	1.212	0.362	0.176	12.066	1.151	1.238	0.257	0.086
RT-LA-VocE	A	✓	17.80±0.03	8.761	1.148	1.205	0.490	0.315	9.934	1.112	1.115	0.376	0.172	10.585	1.166	1.119	0.293	0.086
RT-AV-GCRN	AV	✓	6.56±0.15	9.713	1.496	1.691	0.769	0.607	10.675	1.209	1.472	0.622	0.416	11.087	1.117	1.262	0.452	0.237
RT-AV-Demucs	AV	✓	4.87±0.13	6.228	1.408	1.464	0.729	0.550	7.363	1.202	1.281	0.591	0.375	8.361	1.108	1.239	0.443	0.222
RT-LA-VocE	AV	✓	20.88±0.04	**4.653**	**1.741**	**2.050**	**0.800**	**0.649**	**5.737**	**1.402**	**1.615**	**0.701**	**0.516**	**6.799**	**1.199**	**1.402**	**0.568**	**0.365**

Table 1: *Comparison between RT-LA-VocE and other speech enhancement methods for different noise conditions. The best results for offline inference (among the non-causal models) are underlined, and the best online results (among the causal models) are highlighted in **bold**. "Proc. latency" denotes the processing latency per frame (40 ms) during inference.*

Audio encoder	Temporal model	Spec. inv. method	MCD ↓	PESQ ↑	ViSQOL ↑	STOI ↑	ESTOI ↑	Alg. latency (ms)	Proc. latency (ms)	# Params (M)
Spec. + lin. layer	Emformer ($l_c = 32$)	Noisy phase	6.342	1.278	1.457	0.589	0.39	55 (40 + 15)	13.65±0.13	96.6
Mel-spec. + lin. layer	Emformer ($l_c = 32$)	Noisy phase	5.962	1.33	1.54	0.651	0.448	55 (40 + 15)	19.07±0.25	96.3
Mel-spec. + lin. layer	Emformer ($l_c = 32$)	Griffin-Lim	6.468	1.236	1.39	0.636	0.446	55 (40 + 15)	30.86±0.31	96.3
Mel-spec. + lin. layer	Emformer ($l_c = 32$)	C-HiFi-GAN	5.906	1.391	1.567	0.694	0.504	55 (40 + 15)	18.81±0.04	110
Causal mel-spec.+ lin. layer	Emformer ($l_c = 32$)	C-HiFi-GAN	6.145	1.302	1.463	0.645	0.456	**40**	19.44±0.04	110
Causal 1D ResNet (25 Hz)	Emformer ($l_c = 32$)	C-HiFi-GAN	5.988	1.261	1.421	0.614	0.436	**40**	19.86±0.04	114
Causal 1D ResNet (100 Hz)	Emformer ($l_c = 32$)	C-HiFi-GAN	5.752	1.397	1.592	0.696	0.510	**40**	20.01±0.04	114
Causal 1D ResNet (100 Hz)	Emformer ($l_c = 64$)	C-HiFi-GAN	**5.737**	**1.402**	**1.615**	**0.701**	**0.516**	**40**	20.88±0.04	114

Table 2: *Ablation on the main architectural components of RT-LA-VocE (noise condition 2). l_c denotes left context length.*

Backbone		Online	MCD ↓	PESQ ↑	ViSQOL ↑	STOI ↑	ESTOI ↑
Temp.	Spec. inv.						
Transformer	HiFi-GAN	✗	5.248	1.64	1.826	0.77	0.606
Transformer	C-HiFi-GAN	✗	5.285	1.633	1.815	0.773	0.602
Emformer	HiFi-GAN	✗	5.696	1.422	1.628	0.710	0.527
Emformer	C-HiFi-GAN	✗	5.737	1.402	1.615	0.701	0.516
Transformer	HiFi-GAN	✓	12.394	1.073	1.052	0.125	0.036
Transformer	C-HiFi-GAN	✓	12.502	1.09	1.035	0.12	0.034
Emformer	HiFi-GAN	✓	6.027	1.299	1.445	0.687	0.502
Emformer	C-HiFi-GAN	✓	**5.737**	**1.402**	**1.615**	**0.701**	**0.516**

Table 3: *Online and offline inference results with causal and non-causal models (noise condition 2).*

Mouth crop. method	Latency (ms)	MCD↓	PESQ ↑	ViSQOL↓	STOI ↑	ESTOI↓
LA-VocE [22]	92.65±2.56	**5.737**	**1.402**	**1.615**	**0.701**	**0.516**
MediaPipe [29]	**7.27±0.07**	5.910	1.380	1.577	0.689	0.503

Table 4: *Quantitative results on different mouth crops (noise condition 2). LA-VocE's crops are smoothed using future frames, while MediaPipe's crops are completely causal.*

methods – Griffin-Lim [37], and noisy phase reconstruction (as in [38]). We find that both methods are slower than *C-HiFi-GAN* and are unable to match its quality and intelligibility. In the bottom half of Table 2, in an attempt to eliminate the algorithm latency introduced by the mel-spectrogram input (15 ms), we further propose three causal audio encoders - a causal mel-spectrogram that leverages the causal padding trick, and two causal 1D ResNets that operate on the raw audio instead (see Section 2.2). We conclude that the 100 Hz ResNet achieves the best results and even outperforms the original mel-spectrogram encoder, while reducing the algorithm latency by 15 ms. Finally, we show that increasing the left context length of the Emformer from 32 to 64 yields considerable improvements in performance while having negligible impact on the latency.

Causal vs. non-causal components. We compare the causal and non-causal versions of the temporal encoder (Emformer and Transformer) and neural vocoder (*C-HiFi-GAN* and HiFi-GAN) using two inference modes in Table 3. In offline inference, the enhanced speech is computed using the entire audio-visual input, as in [22]. In online inference, which is the focus of our work, the enhanced audio is generated frame by frame and can only leverage past inputs, as in Figure 1. As expected, we observe that the Transformer and HiFi-GAN surpass their causal counterparts in offline inference, but underperform when generating audio online, since they can no longer leverage future information. In contrast, the Emformer and *C-HiFi-GAN* achieve the same performance in both inference modes since they leverage only past frames, and achieve the best online inference results by a wide margin. This highlights the need to design causal modules, rather than naïvely leveraging the non-causal components presented in [22] for real-time enhancement.

4.3. End-to-end real-time enhancement

Low-latency mouth cropping. In the experiments presented above, we obtain the mouth crops following the pipeline presented in LA-VocE [22], which has an average latency of 92.65 ms. In Table 4, we explore whether our trained models can produce high-quality speech using mouth crops generated by a causal low-latency (7.3 ms per frame) mouth cropping pipeline based on MediaPipe [29]. We find that the performance difference between these two methods is marginal (roughly 2 %), which shows that RT-LA-VocE can effectively adapt to the videos produced by MediaPipe without a significant loss in performance.

Total processing latency. Given RT-LA-VocE's and MediaPipe's latencies per frame (20.88 ms and 7.27 ms, respectively), our end-to-end pipeline has a total processing latency of 28.15 ms per frame. This is considerably lower than the time it takes to obtain the next frame (*i.e.*, 40 ms), meaning that we can perform real-time enhancement without accumulating delays at each time-step.

5. Conclusion

In this paper, we present RT-LA-VocE, a real-time audio-visual speech enhancement model that adapts the original LA-VocE [22] for causal inference. Our proposed architecture sets a new state-of-the-art for real-time AVSE, and is competitive with the original LA-VocE during offline inference, despite not relying on future information. Furthermore, our findings show that RT-LA-VocE can be leveraged via a server-side GPU to enhance noisy speech in live video streams with minimal processing latency (< 30 ms).

6. References

[1] J. Benesty, S. Makino, *et al.*, *Speech enhancement*. Springer Science & Business Media, 2006.

[2] A. Défossez, G. Synnaeve, *et al.*, "Real time speech enhancement in the waveform domain," in *Interspeech*, ISCA, 2020.

[3] A. Pandey and D. Wang, "TCNN: temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP*, IEEE, 2019.

[4] M. Thakker, S. E. Eskimez, *et al.*, "Fast real-time personalized speech enhancement: End-to-end enhancement network (e3net) and knowledge distillation," in *Interspeech*, ISCA, 2022.

[5] T. Gao, J. Du, *et al.*, "Improving deep neural network based speech enhancement in low SNR environments," in *LVA/ICA*, ser. Lecture Notes in Computer Science, Springer, 2015.

[6] L. Birnie, P. N. Samarasinghe, *et al.*, "Noise RETF estimation and removal for low SNR speech enhancement," in *MLSP*, IEEE, 2021.

[7] 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.

[8] K. A. Karl, J. V. Peluchette, *et al.*, "Virtual work meetings during the covid-19 pandemic: The good, bad, and ugly," *Small Group Research*, 2022.

[9] S. Petridis, T. Stafylakis, *et al.*, "End-to-end audiovisual speech recognition," in *ICASSP*, IEEE, 2018.

[10] A. Haliassos, P. Ma, *et al.*, "Jointly learning visual and auditory speech representations from raw data," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023.

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

[12] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in *CVPR*, IEEE, 2021.

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

[14] W.-N. Hsu, T. Remez, *et al.*, "Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration," in *CVPR*, IEEE, 2023.

[15] M. Gogate, K. Dashtipour, *et al.*, "Cochleanet: A robust language-independent audio-visual model for real-time speech enhancement," *Inf. Fusion*, 2020.

[16] M. Gogate, K. Dashtipour, *et al.*, "Towards robust real-time audio-visual speech enhancement," *arXiv*, 2021.

[17] M. Gogate, K. Dashtipour, *et al.*, "Towards real-time privacy-preserving audio-visual speech enhancement," in *SPSC*, ISCA, 2022.

[18] 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*, 2006.

[19] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, 2015.

[20] Z. Zhu, H. Yang, *et al.*, "Real-time audio-visual end-to-end speech enhancement," in *ICASSP*, 2023.

[21] J. F. Montesinos, V. S. Kadandale, *et al.*, "Vovit: Low latency graph-based audio-visual voice separation transformer," in *ECCV*, Springer, 2022.

[22] R. Mira, B. Xu, *et al.*, "LA-VocE: Low-SNR audio-visual speech enhancement using neural vocoders," in *ICASSP*, 2023.

[23] J. Kong, J. Kim, *et al.*, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[24] Y. Shi, Y. Wang, *et al.*, "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in *ICASSP*, IEEE, 2021.

[25] A. Vaswani, N. Shazeer, *et al.*, "Attention is all you need," in *NeurIPS*, Curran Associates, Inc., 2017.

[26] A. van den Oord, S. Dieleman, *et al.*, "Wavenet: A generative model for raw audio," in *Speech Synthesis Workshop*, ISCA, 2016.

[27] A. Ephrat, I. Mosseri, *et al.*, "Looking to listen at the cocktail party," *ACM Transactions on Graphics (TOG)*, 2018.

[28] C. K. Reddy, V. Gopal, *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *INTERSPEECH*, 2020.

[29] C. Lugaresi, J. Tang, *et al.*, "Mediapipe: A framework for building perception pipelines," *CoRR*, 2019.

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

[31] 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*, 2001.

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

[33] C. H. Taal, R. C. Hendriks, *et al.*, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Speech Audio Process.*, 2011.

[34] 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.*, 2016.

[35] 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.*, 2020.

[36] A. Défossez, N. Usunier, *et al.*, "Music source separation in the waveform domain," *arXiv*, 2019.

[37] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," in *ICASSP*, IEEE, 1983.

[38] A. Gabbay, A. Shamir, *et al.*, "Visual speech enhancement," in *Interspeech*, ISCA, 2018.