# VSEGAN: VISUAL SPEECH ENHANCEMENT GENERATIVE ADVERSARIAL NETWORK

*Xinmeng Xu*[1,2], *Yang Wang*[1], *Dongxiang Xu*[1], *Yiyuan Peng*[1], *Cong Zhang*[1], *Jie Jia*[1], *Binbin Chen*[1]

[1]vivo AI Lab, P.R. China
[2]E.E. Engineering, Trinity College Dublin, Ireland

## ABSTRACT

Speech enhancement is an essential task of improving speech quality in noise scenario. Several state-of-the-art approaches have introduced visual information for speech enhancement, since the visual aspect of speech is essentially unaffected by acoustic environment. This paper proposes a novel framework that involves visual information for speech enhancement, by incorporating a Generative Adversarial Network (GAN). In particular, the proposed visual speech enhancement GAN consists of two networks trained in adversarial manner, i) a generator that adopts multi-layer feature fusion convolution network to enhance input noisy speech, and ii) a discriminator that attempts to minimize the discrepancy between the distributions of the clean speech signal and enhanced speech signal. Experiment results demonstrated superior performance of the proposed model against several state-of-the-art models.

***Index Terms***— speech enhancement, visual information, multi-layer feature fusion convolution network, generative adversarial network

## 1. INTRODUCTION

Speech processing systems are used in a wide variety of applications such as speech recognition, speech coding, and hearing aids. These systems have best performance under the condition that noise interference are absent. Consequently, speech enhancement is essential to improve the performance of these systems in noisy background [1]. Speech enhancement is a kind of algorithm that can be used to improve the quality and intelligibility of noisy speech, decrease the hearing fatigue, and improve the performance of many speech processing systems.

Conventional speech enhancement algorithms are mainly based on signal processing techniques, e.g., by using speech signal characteristics of a known speaker, which include spectral subtraction [2], signal subspace [3], Wiener filter [4], and model-based statistical algorithms [5]. Various deep learning networks architectures, such as fully connected network, Convolution Neural Networks (CNNs), Recurrent Neural Networks (RNNs), have been demonstrated to notably improve speech enhancement capabilities than that of conventional approaches. Although deep learning approaches make noisy speech signal more audible, there are some remaining deficiencies in restoring intelligibility.

Speech enhancement is inherently multimodal, where visual cues help to understand speech better. The correlation between the visible proprieties of articulatory organs, e.g., lips, teeth, tongue, and speech reception has been previously shown in numerous behavioural studies [6]. Similarly, a large number of previous works have been developed for visual speech enhancement, which based on signal processing techniques and machine learning algorithms [7]. Not surprisingly, visual speech enhancement has been recently addressed in the framework of DNNs, a fully connected network was used to jointly process audio and visual inputs to perform speech enhancement [8]. The fully connected architecture cannot effectively process visual information, which caused the audio-visual speech enhancement system slightly better than its audio-only speech enhancement counterpart. In addition, there is a model which feed the video frames into a trained speech generation network, and predict clean speech from noisy input [9], which has shown more obvious improvement when compared with the previous approaches.

The Generative Adversarial Network (GAN) consists of a generator network and a discriminator network that play a min-max game between each other, and GAN have been explored for speech enhancement, SEGAN [10] is the first approach to apply GAN to speech enhancement model. This paper proposes a Visual Speech Enhancement Generative Adversarial Network (VSEGAN) that enhances noisy speech using visual information under GAN architecture.

The rest of article is organized as follows: Section 2 presents the proposed method in detail. Section 3 introduces the experimental setup. Experiment results are discussed in Section 4, and a conclusion is summarized in Section 5.

## 2. MODEL ARCHITECTURE

### 2.1. Generative Adversarial Network

GAN is comprised of generator (G) and discriminator (D). The function of G is to map a noisy vector $\mathbf{x}$ from a given prior distribution $\mathcal{X}$ to an output sample $\mathbf{y}$ from the distribution $\mathcal{Y}$ of training data. D is a binary classifier network, which

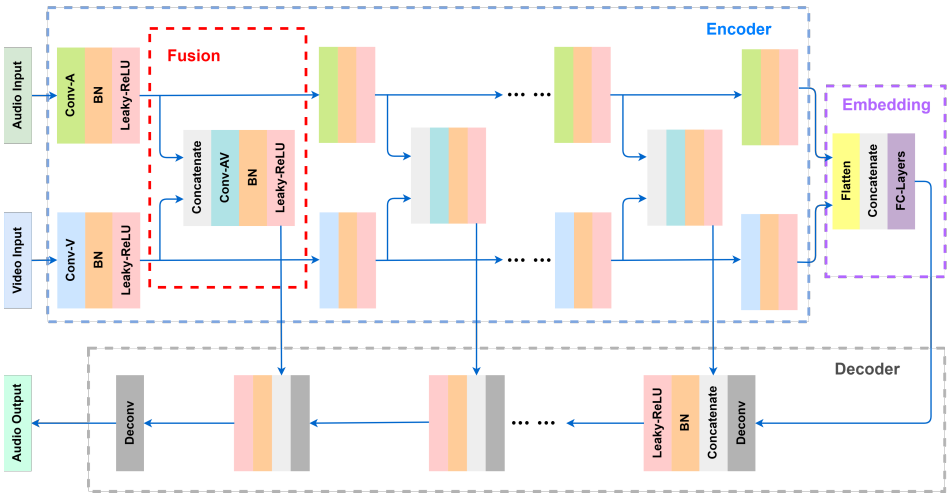


**Fig. 1**. Network architecture of generator. Conv-A, Conv-V, Conv-AV, BN, and Deconv denote convolution of audio encoder, convolution of video encoder, convolution of audio-visual fusion, batch normalization, and transposed convolution.

determines whether its input is real or fake. The generated samples coming from $\mathcal{Y}$, are classified as real, whereas the samples coming from G, are classified as fake. The learning process can be regarded as a minimax game between G and D, and can be expressed by:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{y}}(\mathbf{y})}[\log(D(\mathbf{y}))] + \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}(\mathbf{x})}[\log(1-D(G(\mathbf{x})))] \quad (1)$$

Training procedure for GAN can be concluded the repetition of following three steps:

Step 1: D back-props a batch of real samples $\mathbf{y}$.

Step 2: Freeze the parameters of G, and D back-props a batch of fake samples that generated from G.

Step 3: Freeze the parameters of D, and G back-props to make D misclassify.

The regression task generally works with a conditioned version of GAN [11], in which some extra information, involve in a vector $\mathbf{y}_c$, is provided along with the noisy vector $\mathbf{x}$ at the input of G. In that case, the cost function of D is expressed as following:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{y},\mathbf{y}_c\sim p_{\mathbf{y}}(\mathbf{y},\mathbf{y}_c)}[\log(D(\mathbf{y},\mathbf{y}_c))] + \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}(\mathbf{x}),\mathbf{y}_c\sim p_{\mathbf{y}}(\mathbf{y}_c)}[\log(1-D(G(\mathbf{x},\mathbf{y}_c),\mathbf{y}_c))] \quad (2)$$

However, Eq. (2) are suffered from vanishing gradients due to the sigmoid cross-entropy loss function [12]. To tackle this problem, least-squares GAN approach [13] substitutes cross-entropy loss to the mean-squares function with binary coding, as given in Eq. (3) and Eq. (4).

$$\max_D V(D) = \frac{1}{2}\mathbb{E}_{\mathbf{y},\mathbf{y}_c\sim p_{\mathbf{y}}(\mathbf{y},\mathbf{y}_c)}[\log(D(\mathbf{y},\mathbf{y}_c)-1)^2] + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}(\mathbf{x}),\mathbf{y}_c\sim p_{\mathbf{y}}(\mathbf{y}_c)}[\log(1-D(G(\mathbf{x},\mathbf{y}_c),\mathbf{y}_c))^2] \quad (3)$$

$$\min_G V(G) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}(\mathbf{x}),\mathbf{y}_c\sim p_{\mathbf{y}}(\mathbf{y}_c)}[\log(D(G(\mathbf{x},\mathbf{y}_c),\mathbf{y}_c)-1)^2] \quad (4)$$

### 2.2. Visual Speech Enhancement GAN

The G network of VSEGAN performs enhancement, where its inputs are noisy speech $\tilde{\mathbf{y}}$ and video frames $\mathbf{v}$, and its output is the enhanced speech $\widehat{\mathbf{y}} = G(\tilde{\mathbf{y}}, \mathbf{v})$. The G network follows an encoder-decoder scheme, and consist of encoder part, fusion part, embedding part, and decoder part. The architecture of G network is shown in Figure 1.

Encoder part of G network involves audio encoder and video encoder. The audio encoder is designed as a CNN taking spectrogram as input, and each layer of an audio encoder is followed by strided convolutional layer, batch normalization, and Leaky-ReLU for non-linearity. The video encoder is used to process the input face embedding through a number of max-pooling convolutional layers followed by batch normalization, and Leaky-ReLU. In the G network, the dimension of visual feature vector after convolution layer has to be the same as the corresponding audio feature vector, since both vectors take at every encoder layer is through a fusion part in encoding stage. The audio decoder is reversed in the audio encoder part by deconvolutions, followed again by batch normalization and Leaky-ReLU.

Fusion part designates a merged dimension to implement fusion, and the audio and video streams take the concatenation operation and are through several strided convolution

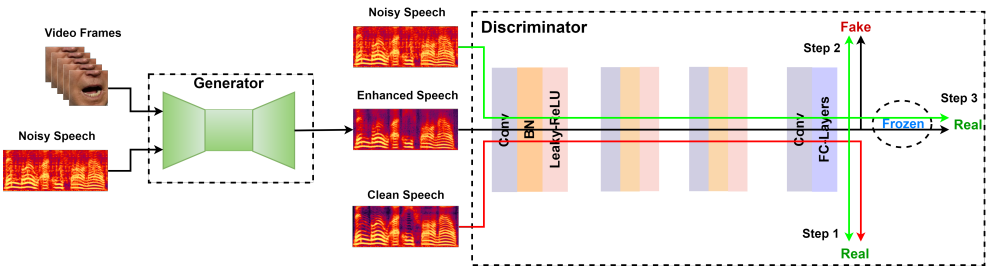


**Fig. 2**. Network architecture of discriminator, and GAN training procedure.

**Table 1**. Detailed architecture of the VSEGAN generator encoders. Conv1 denotes the first convolution layer of the VSEGAN generator encoder part.

| | Conv1 | Conv2 | Conv3 | Conv4 | Conv5 | Conv6 | Conv7 | Conv8 | Conv9 | Conv10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Num Filters | 64 | 64 | 128 | 128 | 256 | 256 | 512 | 512 | 1024 | 1024 |
| Filter Size | (5, 5) | (4, 4) | (4, 4) | (4, 4) | (2, 2) | (2, 2) | (2, 2) | (2, 2) | (2, 2) | (2, 2) |
| Stride(audio) | (2, 2) | (1, 1) | (2, 2) | (1, 1) | (2, 1) | (1, 1) | (2, 1) | (1, 1) | (1, 5) | (1, 1) |
| MaxPool(video) | (2, 4) | (1, 2) | (2, 2) | (1, 1) | (2, 1) | (1, 1) | (2, 1) | (1, 1) | (1, 5) | (1, 1) |

layer, followed by batch normalization, and Leaky-ReLU. Embedding part consists of three parts: 1) flatten audio and visual steams, 2) concatenate flattened audio and visual streams together, 3) feed concatenated feature vector into several fully-connected layers. The output of fusion part in each layer is fed to the corresponding decoder layer. Embedding part is a bottleneck, which applied deeper feature fusion strategy, but with a larger computation expense. The architecture of G network avoids that many low level details could be lost to reconstruct the speech waveform properly, if all information are forced to flow through the compression bottleneck.

The D network of VSEGAN has the same structure with SERGAN [14], as shown in Figure 2. The D can be seen as a kind of loss function, which transmits the classified information (real or fake) to G, i.e., G can predict waveform towards the realistic distribution, and getting rid of the noisy signals labeled to be fake. In addition, previous approaches [15] demonstrated that using $L_1$ norm as an additional component is beneficial to the loss of G, and $L_1$ norm which performs better than $L_2$ norm to minimize the distance between enhanced speech and target speech [16]. Therefore, the G loss is modified as:

$$\min_G V(G) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}(\mathbf{x}),\tilde{\mathbf{y}}\sim p_{\mathbf{y}}(\tilde{\mathbf{y}})}[(D(G(\mathbf{x},(\mathbf{v},\tilde{\mathbf{y}})),\tilde{\mathbf{y}}) - 1)^2] + \lambda||G(\mathbf{x},(\mathbf{v},\tilde{\mathbf{y}})) - \mathbf{y}||_1 \quad (5)$$

where $\lambda$ is a hyper-parameter to control the magnitude of the $L_1$ norm.

## 3. EXPERIMENT SETUP

### 3.1. Datasets

The model is trained on two datasets: the first is the GRID [17] which consist of video recordings where 18 male speakers and 16 female speakers pronounce 1000 sentences each; the second is TCD-TIMIT [18], which consist of 32 male speakers and 30 female speakers with around 200 videos each.

The noise signals are collected from the real world and categorized into 12 types: room, car, instrument, engine, train, talker speaking, air-brake, water, street, mic-noise, ring-bell, and music. At every iteration of training, a random attenuation of the noise interference in the range of [-15, 0] dB is applied as a data augmentation scheme. This augmentation was done to make the network robust against various SNRs.

### 3.2. Training and Network Parameters

The video representation is extracted from input video and is resampled to 25 frames per seconds. Each video is divided into non-overlapping segments of 5 consecutive frames. The audio representation is the transformed magnitude spectrograms in the log Mel-domain with 80 Mel frequency bands from 0 to 8 kHz, using a Hanning window of length 640 bins (40 milliseconds), and hop size of 160 bins (10 milliseconds). The whole spectrograms are sliced into pieces of duration of 200 milliseconds corresponding to the length of 5 video frames.

The proposed VSEGAN has 10 convolutional layers for each encoder and decoder of generator, and the details of au-

**Table 2**. Performance of trained networks

| Test SNR | -5 dB | | 0 dB | |
|---|---|---|---|---|
| Evaluation Metrics | STOI | PESQ | STOI | PESQ |
| Noisy | 51.4 | 1.03 | 62.6 | 1.24 |
| SEGAN | 63.4 | 1.97 | 77.3 | 2.21 |
| Baseline | 81.3 | 2.35 | 87.9 | 2.94 |
| VSEGAN | **86.8** | **2.88** | **89.8** | **3.10** |

**Table 3**. Performance comparison of VSEGAN with state-of-the-art result on GRID

| Test SNR | -5 dB | 0 dB | -5 dB | 0 dB |
|---|---|---|---|---|
| Evaluation Metrics | PESQ | | STOI(%) | |
| L2L | 2.61 | 2.92 | 85.89 | 88.96 |
| OVA | 2.69 | 3.00 | 86.17 | 89.75 |
| AV(SE)$^2$ | - | 2.98 | 86.06 | 89.44 |
| VSEGAN | 2.88 | 3.10 | 86.84 | 89.83 |

dio and visual encoders are described in Table 1, and a Conv-A or a Conv-V in Figure 1 comprise of two convolution layers in Table 1.

The model is trained with ADAM optimizer for 70 epochs with learning rate of $10^{-4}$, and batch size of 8, and the hyper parameter $\lambda$ of loss function in Eq. (5) is set to 100.

## 4. RESULTS

The performance of VSEGAN is evaluated with the following metrics: Perceptual Evaluation of Speech Quality (PESQ), and Short Term Objective Intelligibility (STOI). In addition, there are three networks have trained for comparison:

- **SEGAN** [10]: An audio-only speech enhancement generative adversarial network.
- **Baseline** [19]: A baseline work of visual speech enhancement.
- **VSEGAN**: the proposed model, visual speech enhancement generative adversarial network.

Table 2 demonstrates the improvement performance of network, as a new component is added to the architecture, such as visual information, multi-layer feature fusion strategy, and finally GAN model. The VSEGAN outperforms SEGAN, which is an evidence that visual information significantly improves the performance of speech enhancement system. What is more, the comparison between VSEGAN and baseline illustrates that GAN model for visual speech enhancement is more robust than G-only model. Hence the performance improvement from SEGAN to VSEGAN is primarily for two reason: 1) using visual information, and 2) using GAN model. Figure 3 shows the visualization of baseline system enhancement, Generator-only enhancement, and VSEGAN enhancement, which most obvious details of spectrum distinction are framed by dotted box.[1]

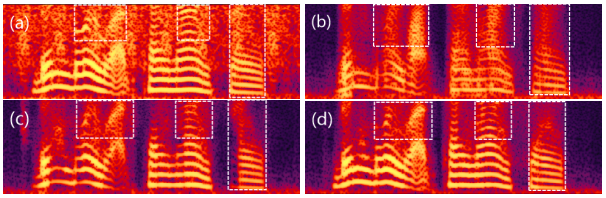


**Fig. 3**. Example of input and enhanced spectra from an example speech utterance. (a) Noisy speech under the condition of noise at 0 dB. (b) Enhanced speech generated by baseline work. (c) Enhanced speech generated by Generator. (d) Enhanced speech generated by VSEGAN.

For further investigating the superiority of proposed method, the performance of VSEGAN has also compared to the following recent audio-visual speech enhancement approaches on GRID dataset:

- **Looking-to-Listen model** [20]: A speaker independent audio-visual speech separation model.
- **Online Visual Augmented (OVA) model** [21]: A late fusion based visual speech enhancement model, which involves the audio-based component, visual-based component and the augmentation component.
- **AV(SE)$^2$ model** [22]: An audio-visual squeeze-excite speech enhancement model.

Table 3 shows that the VSEGAN produces state-of-the-art results in terms of PESQ and STOI score by comparing against four recent proposed methods that use DNNs to perform end-to-end visual speech enhancement. Results for competing methods are taken from the corresponding papers and the missing entries in the table indicate that the metric is not reported in the reference paper. Although the competing results are for reference only, the VSEGAN has better performance than state-of-the-art results on the GRID dataset.

## 5. CONCLUSIONS

This paper proposed an end-to-end visual speech enhancement method has been implemented within the generative adversarial framework. The model adopts multi-layer feature fusion convolution network structure, which provides a better training behavior, as the gradient can flow deeper through the whole structure. According to the experiment results, the performance of speech enhancement system has significantly improves by involving of visual information, and visual speech enhancement using GAN performs better quality of enhanced speech than several state-of-the-art models.

[1]Speech samples are available at: `https://XinmengXu.github.io/AVSE/VSEGAN`

## 6. REFERENCES


[1] Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[2] Rainer Martin, "Spectral subtraction based on minimum statistics," *power*, vol. 6, no. 8, 1994.

[3] Yariv Ephraim and Harry L Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on speech and audio processing*, vol. 3, no. 4, pp. 251–266, 1995.

[4] Jae Lim and Alan Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[5] Markos Dendrinos, Stelios Bakamidis, and George Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991.

[6] William H Sumby and Irwin Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[7] Wenwu Wang, Darren Cosker, Yulia Hicks, S Saneit, and Jonathon Chambers, "Video assisted speech source separation," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* IEEE, 2005, vol. 5, pp. v–425.

[8] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Jen-Chun Lin, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang, "Audio-visual speech enhancement using deep neural networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[9] Aviv Gabbay, Ariel Ephrat, Tavi Halperin, and Shmuel Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3051–3055.

[10] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Interspeech 2017*, 2017, pp. 3642–3646.

[11] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *Computer ence*, pp. 2672–2680, 2014.

[12] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 11 2016.

[13] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[14] Deepak Baby and Sarah Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 106–110.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[16] Ashutosh Pandey and Deliang Wang, "On adversarial training and loss functions for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5414–5418.

[17] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[18] Naomi Harte and Eoin Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[19] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg, "Visual speech enhancement," *Interspeech*, pp. 1170–1174, 2018.

[20] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, 2018.

[21] Wupeng Wang, Chao Xing, Dong Wang, Xiao Chen, and Fengyu Sun, "A robust audio-visual speech enhancement model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7529–7533.

[22] Michael L Iuzzolino and Kazuhito Koishida, "AV(SE)$^2$: Audio-visual squeeze-excite speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7539–7543.