# A Diffusion-BASED AUDIO-VISUAL SPEECH ENHANCEMENT APPROACH FOR THE 3RD COG-MHEAR AVSE CHALLENGE

*Chia-Wei Chen*[1], *Jun-Cheng Chen*[2], *Yu Tsao*[2], *Shao-Yi Chien*[1]

[1]National Taiwan University, Taiwan [2]Academia Sinica, Taiwan

## Abstract

In this work, we proposed a diffusion-based audio-visual speech enhancement (DAVSE) and present the preliminary experiment results. Specifically, we incorporate audio and visual information into a score-based diffusion model [1]. Diffusion models can effectively produce high-quality and intelligible speech by iteratively refining the signal. The visual cues can further assist in distinguishing between different speakers and leading to better generalization and performance in speech enhancement.

## System Description

Our proposed DAVSE system is shown in Figure 1. Noisy speech is converted into spectrogram as the input of the diffusion model. The lips movement information cropped from the video is first processed by a pretrained visual encoder. Then the embedded visual features will be fed into cross-attention modules to more effectively couple the audio-visual features. The backbone model in the diffusion process is NCSN++[2], which is the same backbone model in SGMSE[1]. The number of trainable parameters of the model is 57.7 M. We trained our system on one RTX 3080 (12GB memory) for 68 epochs, which takes about six days (about 2 hours per epoch).

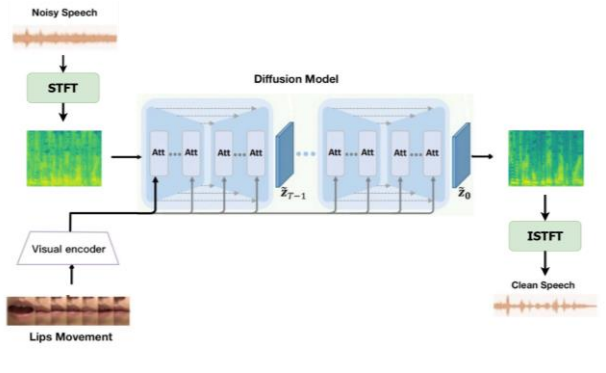


Figure 1: Schematic diagram of DAVSE.

## Experimental Results

We use the LRS3 dataset provided by AVSE challenge (AVSEC) to train our model. The dataset contains 34517 videos in training set, 3306 videos in development set and 2400 videos in evaluation set. The overview results on the evaluation set are shown in Table 1. It is evident that our method significantly improves the quality and intelligibility of noisy speech.

Table 1: Comparison of the evaluation scores of the enhanced speech on the evaluation set.

| Methods | Track | PESQ | STOI | SISDR |
|---|---|---|---|---|
| Noisy | 1 | 1.46 | 0.61 | -5.49 |
| DAVSE | 1 | **1.97** | **0.70** | **1.89** |

## References

[1] J. Richter, S. Welker, J. -M. Lemercier, B. Lay and T. Gerkmann, "Speech Enhancement and Dereverberation With Diffusion-Based Generative Models," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2351-2364, 2023, doi: 10.1109/TASLP.2023.3285241.

[2] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B.Poole, "Score-based generative modeling through stochastic differential equations," in Proc. Int. Conf. Learn. Representations, 2021.