國立臺灣大學電機資訊學院電子工程學研究所

碩士論文

Graduate Institute of Electronics Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

DAVSE: 基於擴散模型的生成式影像結合語音增強方法

DAVSE: A Diffusion-Based Generative Approach for Audio-Visual Speech Enhancement

陳嘉偉

Chia-Wei Chen

指導教授: 簡韶逸 博士

Advisor: Shao-Yi Chien Ph.D.

中華民國 114 年 8 月

August, 2025

doi:10.6342/NTU202502501

Acknowledgements

在本論文完成之際，謹向所有在我研究旅程中給予支持與協助的人致上最深的感謝。

首先，衷心感謝我的指導教授簡韶逸博士，在我攻讀碩士期間給予我無私的指導與鼓勵。一開始因為我做的題目較為冷門，偶爾會感到焦慮，但簡老師不僅非常支持我的研究方向，更時常在我遇到瓶頸時為我點明可行的方向，讓我能夠一步步穩健前行。老師對於學術的熱情與開放態度，深深影響了我對研究的看法，成為我持續努力的動力。

同時，我亦感謝中研院的曹昱博士，在研究合作與討論中提供了許多關鍵性的指導與技術建議，並且大力支持我發表論文。每次與曹博士開會時都會被他的學識淵博震驚，曹博士的專業知識與獨到的見解讓我對於語音增強技術有更深入的理解，也讓我在研究上更加有自信。

此外，也要感謝實驗室的所有同學們，感謝你們在日常生活與研究上給予我陪伴與支持。大家一起互相幫忙解決研究上的難題，以及在研究之餘的日常閒聊，都是我最珍貴的回憶。

再次感謝所有曾經幫助過我的人，是你們的存在，讓這段碩士旅程豐富且充實。

doi:10.6342/NTU202502501

doi:10.6342/NTU202502501

摘要

近年來，語音視覺語音增強（Audio-Visual Speech Enhancement, AVSE）因其能夠在嘈雜環境中提升語音可懂度與品質，受到廣泛關注。儘管去噪效能已有顯著進展，AVSE 系統仍面臨兩項主要挑戰：（1）判別式方法可能引入不自然的語音失真，抵消降噪帶來的效益；（2）視覺訊號的整合往往伴隨額外的運算成本。

本論文提出一種基於擴散模型的創新方法，旨在解決上述挑戰。我們的系統採用一個基於分數的擴散模型來學習乾淨語音資料的先驗分佈。透過這一先驗知識，系統能從偏離學習分佈的嘈雜或混響輸入中推斷出乾淨語音。此外，音訊與視覺輸入透過交叉注意力模組整合至條件噪聲分數網路中，並未增加額外的計算成本。

實驗結果顯示，所提出的 DAVSE 系統在提升語音品質與減少生成性瑕疵（如語音混淆）方面，相較於僅使用音訊的語音增強系統有明顯優勢。此外，實驗也證實交叉注意力模組能有效地融合音訊與視覺資訊。

關鍵字：影像結合語音增強、擴散模型、自然語言處理、深度學習

doi:10.6342/NTU202502501

doi:10.6342/NTU202502501

Abstract

In recent years, audio-visual speech enhancement (AVSE) has attracted considerable attention for its ability to improve speech intelligibility and quality in noisy environments. Despite advances in denoising performance, two major challenges remain in AVSE systems: (1) discriminative approaches can introduce unpleasant speech distortions that may negate the benefits of noise reduction, and (2) integrating visual input often leads to increased processing costs.

This thesis presents a novel diffusion model-based approach to address these challenges. Our system utilizes a score-based diffusion model to learn the prior distribution of clean speech data. This prior knowledge enables the system to infer clean speech from noisy or reverberant input signals that deviate from the learned distribution. In addition, audio and visual inputs are integrated into the noise conditional score network through cross-attention modules, without incurring additional computational costs.

Experimental evaluations demonstrate that the proposed DAVSE system significantly

doi:10.6342/NTU202502501

improves speech quality and reduces generative artifacts, such as phonetic confusions, compared to audio-only SE systems. Furthermore, the results confirm the effectiveness of cross-attention modules in seamlessly incorporating audio and visual information.

Keywords: Audio-Visual Speech Enhancement, Diffusion Models, Natural Language Processing, Deep Learning

doi:10.6342/NTU202502501

Contents

doi:10.6342/NTU202502501

doi:10.6342/NTU202502501

doi:10.6342/NTU202502501

doi:10.6342/NTU202502501

List of Figures

doi:10.6342/NTU202502501

doi:10.6342/NTU202502501

List of Tables

doi:10.6342/NTU202502501

doi:10.6342/NTU202502501

Chapter 1 Introduction

1.1 Audio-Visual Speech Enhancement (AVSE)

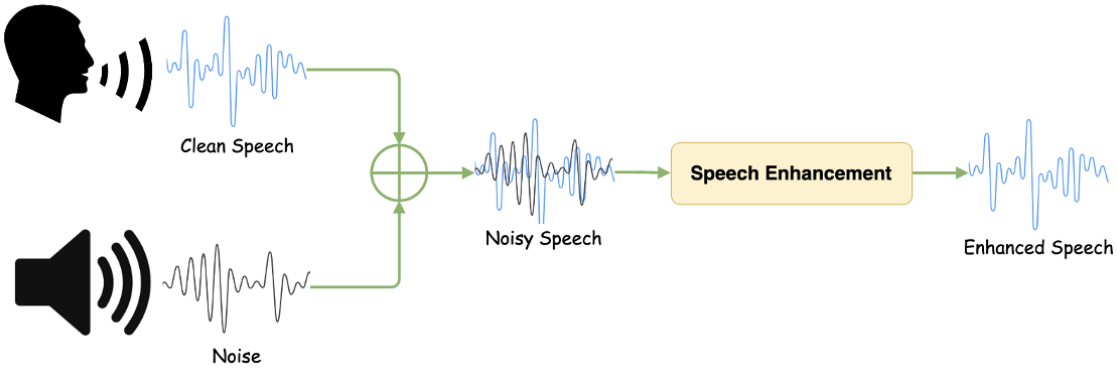
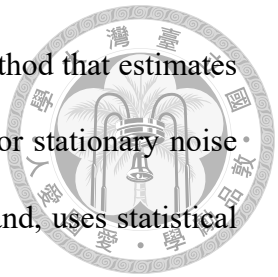


Figure 1.1: Illustration of the speech enhancement task.
A noisy speech signal is processed by a speech enhancement model to suppress background noise and recover the clean speech signal.

Speech Enhancement (SE) is a fundamental technique designed to improve the quality and intelligibility of speech signals, which is crucial for various applications such as telecommunication, hearing aids, voice-controlled systems, and noise reduction in recordings [7]. An overview of the speech enhancement task is illustrated in Figure 1.1. The primary goals of SE are to reduce noise, improve intelligibility, and enhance the overall quality of the speech, making it clearer and more pleasant for listeners. SE aims to remove or minimize the effects of background noise, reverberation, and other distortions that often make communication difficult [12]. Standard methods for SE include traditional approaches like spectral subtraction, Wiener filtering, and Minimum Mean Square

doi:10.6342/NTU202502501



Error (MMSE) estimation [8, 26]. Spectral subtraction is a simple method that estimates and subtracts noise from the noisy speech spectrum. It is effective for stationary noise but is prone to introducing artifacts. Wiener filtering, on the other hand, uses statistical models to minimize the error between the estimated and actual signals, striking a balance between noise reduction and maintaining speech quality. Another approach, MMSE, employs statistical models to minimize distortion and can be effective given proper noise modeling.

Recently, methods based on deep learning have gained popularity for speech enhancement [42]. As illustrated in Figure 1.2, neural network models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures have demonstrated significant improvements over traditional techniques. The improvement is mainly due to their capacity to learn complex noise patterns and relationships within audio signals [9, 28]. Subspace methods, such as Principal Component Analysis (PCA), are also used for separating speech and noise within lower-dimensional spaces, though they are computationally expensive for real-time use [13]. While traditional methods are more straightforward and less resource-intensive, deep learning-based approaches offer higher enhancement quality but come with demands for large training datasets and computational power [41].

The process of Audio-Visual Speech Enhancement (AVSE), illustrated in Figure 1.3, advances Speech Enhancement (SE) by integrating visual cues such as lip movements, facial expressions, and gestures to enhance speech quality in noisy environments [3]. AVSE can significantly outperform traditional Audio-Only Speech Enhancement (AOSE) methods by leveraging audio and visual information, especially in challenging environments. Visual cues are helpful in disambiguating speech content, making AVSE particularly ben-

doi:10.6342/NTU202502501

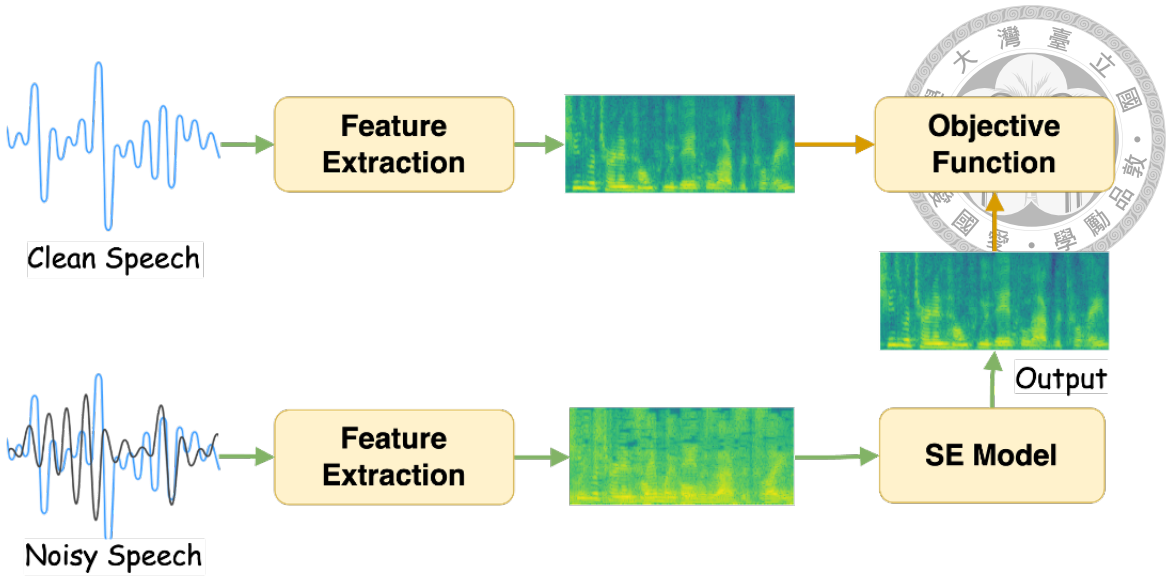


Figure 1.2: Illustration of audio-only speech enhancement (AOSE) system

eficial for applications such as video conferencing, meeting transcription, and assistive technologies for people with hearing impairments [20]. Methods commonly used in AVSE include feature fusion techniques, which integrate visual and audio features either at an early stage (early fusion), at a later decision-making stage (late fusion), or a hybrid of both [2]. Deep learning approaches are also widely adopted, where CNNs are used for extracting visual features from video frames, and RNNs like LSTMs are employed to capture temporal relationships in both modalities [14]. Transformer models have become increasingly popular due to their ability to model long-range dependencies effectively through attention mechanisms, which is crucial in integrating audio-visual information [40].

Generative Adversarial Networks (GANs) are another innovative method used in AVSE, where a generator tries to produce enhanced speech conditioned on visual data, and a discriminator evaluates its quality [31]. This adversarial setup can generate high-quality outputs, but training GANs is notoriously unstable and resource-intensive. [17]. Attention mechanisms are also vital for AVSE systems, allowing models to focus more on relevant features such as lip movements, thereby improving the synchronization and effectiveness of the enhancement [40].

doi:10.6342/NTU202502501

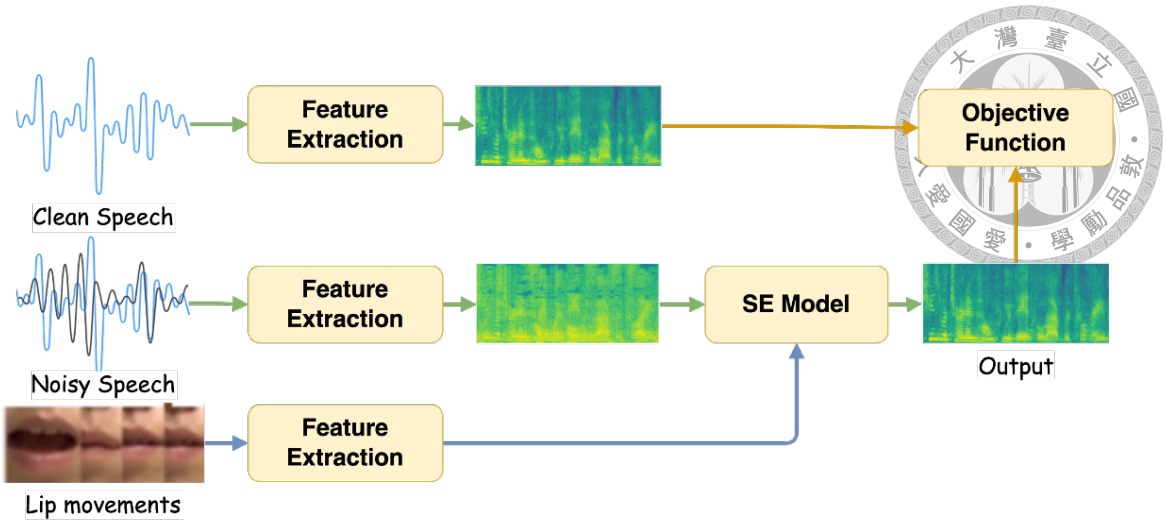


Figure 1.3: Illustration of AVSE system

However, AVSE faces several challenges. One of the primary difficulties is maintaining synchronization between audio and visual streams, as any misalignment can negatively impact performance [3]. Additionally, the quality of visual input is often compromised by occlusion (like when a speaker's mouth is covered) or poor lighting conditions, which reduces the efficacy of the visual information. The variability in speakers' facial features, lip movements, and articulation styles also presents a challenge, as these differences can make it hard for models to generalize well across different individuals [16].

Moreover, the computational complexity of AVSE systems is a significant concern, especially for real-time applications, as integrating visual information alongside audio requires additional processing power. Generalizing unseen environments, where noise, lighting, and facial features vary, is also challenging, as AVSE systems may not constantly adapt well to new conditions. Despite these challenges, AVSE offers a substantial improvement in scenarios where traditional SE struggles, making it a promising approach for enhancing speech quality in real-world noisy settings. Both SE and AVSE benefit from ongoing advances in deep learning, which continue to push the boundaries of what is possible in enhancing speech quality and intelligibility in adverse conditions [41].

doi:10.6342/NTU202502501

1.2 Diffusion Models

A diffusion model is an emerging technique in the realm of generative models, initially gaining prominence for its success in image generation tasks [18]. Recently, diffusion models have found applications in speech enhancement (SE), showing promising results in addressing the challenges associated with noisy or degraded audio signals [32]. The fundamental concept behind diffusion models is to progressively transform a clean input into noise over a series of steps (a diffusion process) and then learn to reverse this process to reconstruct the original input [36]. For SE, this means learning to model how clean speech transforms into noisy speech and then training a system to reverse that degradation, effectively separating clean speech from noise in an iterative manner.

Diffusion models have several benefits when applied to speech enhancement. One significant advantage is their ability to generate high-quality and realistic audio outputs. Unlike traditional SE techniques that might rely on strict assumptions about noise distributions or simplistic transformations, diffusion models take advantage of a probabilistic framework that allows for more flexible and detailed modeling of both speech and noise. This leads to enhanced audio output with fewer artifacts, which often plague traditional SE methods like spectral subtraction or Wiener filtering [8]. The iterative nature of the denoising process in diffusion models contributes to a more refined and gradual removal of noise, thereby maintaining the naturalness and nuances of the original speech signal.

Another key benefit is the robustness of diffusion models when dealing with a wide range of noise types and intensities. By training on diverse datasets, diffusion models can generalize effectively across different acoustic conditions, making them more adaptable to real-world scenarios than traditional methods, which often struggle with non-stationary

doi:10.6342/NTU202502501

or highly unpredictable noise environments [21]. Moreover, diffusion models can be conditioned on auxiliary information, such as speaker identity or contextual features, further improving the quality of the enhanced speech [27].

Diffusion models are also less prone to overfitting than other deep learning-based approaches. Their probabilistic structure allows them to handle uncertainty and variability in noisy conditions more effectively, which helps in providing reliable performance even in unfamiliar environments [36]. Additionally, the advancement in computational power and optimization techniques has made it feasible to deploy diffusion models in practical SE scenarios. However, the computational burden remains higher compared to simpler methods [21].

1.3 Thesis Motivation

In order to overcome the limitations and challenges faced by traditional discriminative methods in speech enhancement, particularly in complex and noisy environments, we built a diffusion-based Audio-Visual Speech Enhancement (DAVSE) system. Traditional AVSE systems, including those based on discriminative deep learning models, are often highly sensitive to the quality of both the audio and visual inputs. In contrast, a diffusion-based AVSE system provides a powerful generative approach that promises to address these challenges by utilizing a fundamentally different paradigm for enhancing speech quality.

One of the key motivations is the robustness offered by diffusion models. Unlike discriminative methods that learn direct mapping from noisy to clean speech, diffusion models learn to progressively refine noisy inputs through a series of probabilistic denois-

doi:10.6342/NTU202502501

ing steps [14]. This iterative refinement process allows the system to effectively handle complex and highly non-linear noise conditions, which are typically challenging for discriminative models. Integrating the visual component allows a diffusion-based AVSE to leverage visual information more robustly, allowing for gradual and accurate restoration of clean speech even when audio cues are heavily degraded. This is particularly important in scenarios where traditional AVSE methods might fail due to low-quality visual inputs, such as mouth occlusion or poor lighting conditions, where a single-step discriminative model might struggle to recover clean speech effectively.

Furthermore, diffusion models have demonstrated a strong capacity for generalization across diverse noise environments, partly due to their ability to effectively model complex distributions and uncertainties. In a diffusion-based AVSE system, including visual cues as an auxiliary source of information allows the model to operate with more context, enhancing robustness in novel noise conditions. Discriminative models often suffer when faced with unseen noise types or speaker variations due to their reliance on fixed mappings. Still, the probabilistic nature of diffusion models, paired with the additional visual information, allows them to adapt more flexibly. This is particularly beneficial in real-world scenarios where noise conditions are highly unpredictable and speakers vary widely in their articulation and facial features.

Additionally, diffusion models offer a high degree of flexibility in incorporating conditioning mechanisms that enhance performance by selectively attending to the most relevant audio and visual features. In our DAVSE system, we leverage cross-attention mechanisms within the attention layers to dynamically emphasize informative visual cues—such as lip movements and facial expressions—throughout the iterative denoising process. This approach effectively mitigates challenges commonly faced by traditional discriminative

doi:10.6342/NTU202502501

methods, which often struggle with accurate time alignment between modalities and incur extra computational costs when handling speaker-specific variations in visual speech patterns. In contrast, the sequential structure of diffusion models enables our system to naturally capture fine-grained temporal alignment and long-range dependencies across audio and visual streams, resulting in more coherent and robust speech enhancement without additional computational burden.

Finally, a diffusion-based AVSE system is well-suited to reduce the artifacts commonly associated with discriminative enhancement methods. Traditional AVSE techniques may introduce unnatural distortions or ”musical noise” due to over-reliance on learned filters and assumptions about noise. In contrast, due to their generative nature, diffusion models work by reconstructing speech progressively, reducing the risk of generating such artifacts and leading to more natural-sounding outputs. This approach aligns well with the goal of producing high-quality, intelligible speech that retains the natural rhythm and tonality of the original speaker, even in challenging conditions.

In summary, the motivation for building a diffusion-based AVSE system lies in its ability to address critical shortcomings of discriminative AVSE methods. These include improving robustness in noise reduction, effectively integrating visual cues to maintain synchronization, generalizing across diverse and challenging environments, and reducing enhancement artifacts. By adopting a generative diffusion approach, we aim to create an AVSE system that offers superior performance in real-world, complex settings, where both audio and visual cues need to be integrated effectively for optimal speech enhancement.

doi:10.6342/NTU202502501

1.4 Contributions

Our contributions to AVSE are centered around three key advancements that address existing challenges while enhancing performance and efficiency.

1. Proposed DAVSE framework for Diffusion-Based AVSE with noise-robust latent visual representation. We developed DAVSE, as shown in 3.1, leveraging diffusion models conditioned on noise-robust latent visual representations. Unlike traditional methods that struggle with degraded visual input, our approach ensures robust enhancement by effectively extracting and utilizing reliable visual features under challenging conditions like occlusions and low lighting. This conditioning allows DAVSE to integrate visual cues seamlessly, significantly improving speech quality in adverse scenarios.

2. Efficient audio-visual feature fusion using cross-attention. We introduced a cross-attention mechanism to efficiently fuse audio and visual features without incurring additional computational costs. Cross-attention dynamically integrates both modalities, allowing the model to focus on relevant information from each. This mechanism improves fusion efficiency compared to conventional methods that often require heavy concatenation or multilayer operations, enabling high-quality AVSE suitable for real-time use while maintaining computational feasibility.

3. We demonstrated the effectiveness of drift supervised loss for enhancing AVSE. This loss function supervises the diffusion model's denoising trajectory, guiding it toward clean speech while minimizing artifacts and distortions. By directly regulating each denoising step, drift supervised loss enhances the stability and naturalness of the enhanced speech, significantly boosting overall quality even in complex, noisy environments.

doi:10.6342/NTU202502501

In summary, our contributions through DAVSE include noise-robust conditioning for enhanced speech quality, efficient audio-visual fusion with cross-attention, and improved model stability with drift-supervised loss. These innovations together set a new benchmark for AVSE, providing effective and efficient enhancement suitable for real-world conditions.

1.5 Thesis Organization

This chapter introduces AVSE and explores the challenges of integrating a user-friendly AVSE system. Next, we summarize the key contributions of this thesis. Chapter 2 reviews relevant works in the field. Chapter 3 details our proposed method for building an AVSE system, emphasizing critical considerations. The experimental results are analyzed in Chapter 4, and the thesis concludes with Chapter 5.

doi:10.6342/NTU202502501

Chapter 2 Related Works

2.1 Audio-Visual Speech Enhancement

Audio-Visual Speech Enhancement (AVSE) has emerged as a promising approach for improving speech quality and intelligibility, especially in noisy environments. Recent research has focused on utilizing deep learning techniques to combine audio and visual cues for enhanced speech quality effectively. Afouras et al. [1] pioneered the use of deep neural networks for AVSE by introducing a model that jointly processed audio spectrograms and visual lip-reading data using a combination of convolutional and recurrent layers. This approach demonstrated that incorporating visual cues can significantly improve the quality of enhanced audio, especially in environments with high background noise. Building on this foundation, Hou et al. [19] introduced AVSEGAN, a generative adversarial network that integrated audio and visual features for robust speech enhancement in challenging acoustic conditions. AVSEGAN employed a generator-discriminator setup, which yielded substantial improvements over previous AVSE approaches regarding speech quality and intelligibility. Chuang et al. [11] contributed to AVSE by introducing an improved lightweight model for AVSE, called improved lite AVSE (ILAVSE), demonstrating significant improvements in efficiency and enhancement quality. The structure of ILAVSE is shown in 2.1. This work emphasized the effectiveness of utilizing visual cues,

doi:10.6342/NTU202502501

particularly the mouth region of interest (ROI), for enhancing speech under noisy conditions. The mouth ROI was shown to play a crucial role in improving the performance of the AVSE model, especially when computational resources are limited. In another recent work, Chern et al. [22] proposed a novel approach for audio-visual speech enhancement and separation using the generalization of the multi-modal self-supervised learning model AV-Hubert [35]. The authors leveraged audio and visual modalities to train embeddings without explicit supervision, resulting in a state-of-the-art (SOTA) performance on the TMSV dataset. This approach effectively utilized the synergies between modalities, providing superior results for both enhancement and separation tasks. This work is the baseline for our research due to its strong performance and novel use of multi-modal self-supervised learning. Another noteworthy contribution came from Gabbay et al. [15],

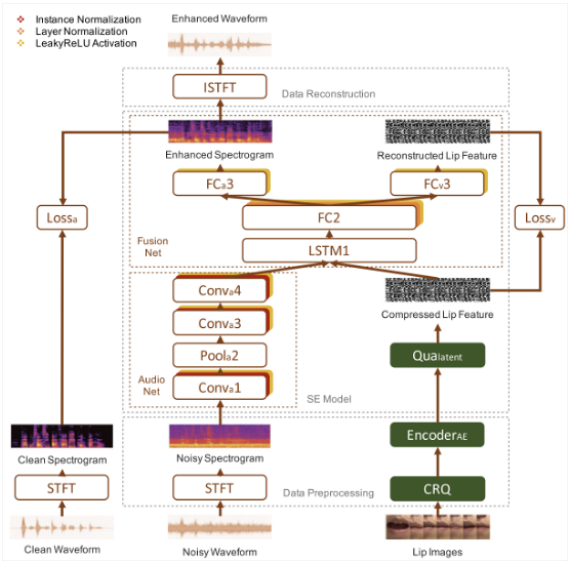
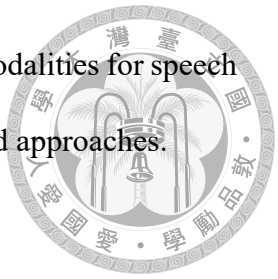


Figure 2.1: The overview of ILAVSE [11]

who explored speech enhancement using only visual inputs, specifically the speaker's lip movements. This work demonstrated that high-quality speech could be reconstructed even under extreme noise conditions when the visual modality was the primary source of information. Michelsanti et al. [30] provided a comprehensive survey of AVSE tech-

doi:10.6342/NTU202502501

niques, detailing the different methods used to fuse visual and audio modalities for speech enhancement and highlighting the effectiveness of deep learning-based approaches.



2.2 Score-based Diffusion Models

Ho et al. [18] proposed Denoising Diffusion Probabilistic Models (DDPMs), which modeled the generative process as a series of denoising steps. This approach formed a foundational methodology for subsequent work involving diffusion models and underscored the value of iterative denoising in generative modeling tasks. Kong et al. [21] extended diffusion models to audio applications with the introduction of DiffWave, a versatile diffusion model tailored for audio synthesis. DiffWave demonstrated the utility of diffusion processes in generating realistic and coherent audio signals, making it highly relevant for audio and speech enhancement tasks. Around the same time, Score-based diffusion models have gained significant attention in recent years, particularly for their application in generative tasks involving iterative noise estimation. These models, originally designed to model data distributions, have shown considerable potential in enhancing the quality of generated signals. Song et al. [37] contributed significantly to this field by introducing score-based generative models using Stochastic Differential Equations (SDEs). The schematic of score-based generative models is shown in 2.2. This work laid the groundwork for training generative models by learning score functions, enabling efficient signal reconstruction and generation. Further advancing these concepts, Chen et al. [10] introduced WaveGrad 2, which used an iterative refinement approach to enhance speech quality through denoising. WaveGrad 2's application of diffusion techniques to speech denoising showcased promising results, suggesting that diffusion models could be effectively adapted to the AVSE domain for producing high-quality, noise-free speech.

doi:10.6342/NTU202502501

In summary, the existing body of work in AVSE has shown that the integration of visual and audio signals using deep learning techniques can significantly improve speech quality, particularly in noisy environments. With their iterative approach to noise estimation, score-based diffusion models present a compelling new method for generating and refining audio signals. The intersection of these fields offers promising opportunities for advancing AVSE, primarily through innovative diffusion-based techniques that can effectively tackle the challenges posed by noisy audio environments.

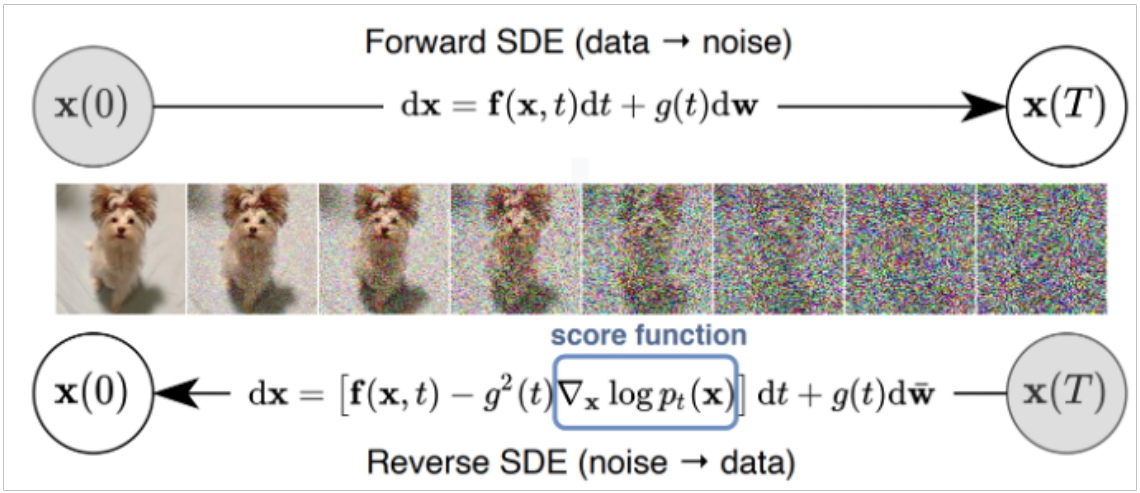


Figure 2.2: The Schematic of score-based generative model

doi:10.6342/NTU202502501

Chapter 3 Proposed Methods

Figure 3.1 shows the overview of our proposed Diffusion-based Audio-Visual Speech Enhancement (DAVSE) system. The system takes advantage of both audio and visual cues to enhance speech quality, particularly in noisy environments. The main components include spectrogram conversion, visual feature extraction, cross-attention modules, and diffusion-based denoising, which work in concert to produce high-quality enhanced speech.

The noisy speech is initially converted into a spectrogram using the Short-Time Fourier Transform (STFT). This process is mathematically represented as follows:

$$Y(t,f) = \sum_{n=-\infty}^{\infty} y[n]w[n-t]e^{-j2\pi fn} \tag{3.1}$$

Where $Y(t,f)$ is the spectrogram, $y[n]$ is the sampled audio signal, $w[n]$ is the windowing function, t is the time index, and f is the frequency index. The spectrogram representation allows the model to operate in the frequency domain, where both magnitude and phase information are crucial for effective speech enhancement. The spectrogram then serves as the input for our diffusion model, which plays a central role in denoising and reconstructing the clean speech signal. By modeling the spectrogram as a probability distribution, the diffusion process iteratively reduces noise through a series of reverse

doi:10.6342/NTU202502501

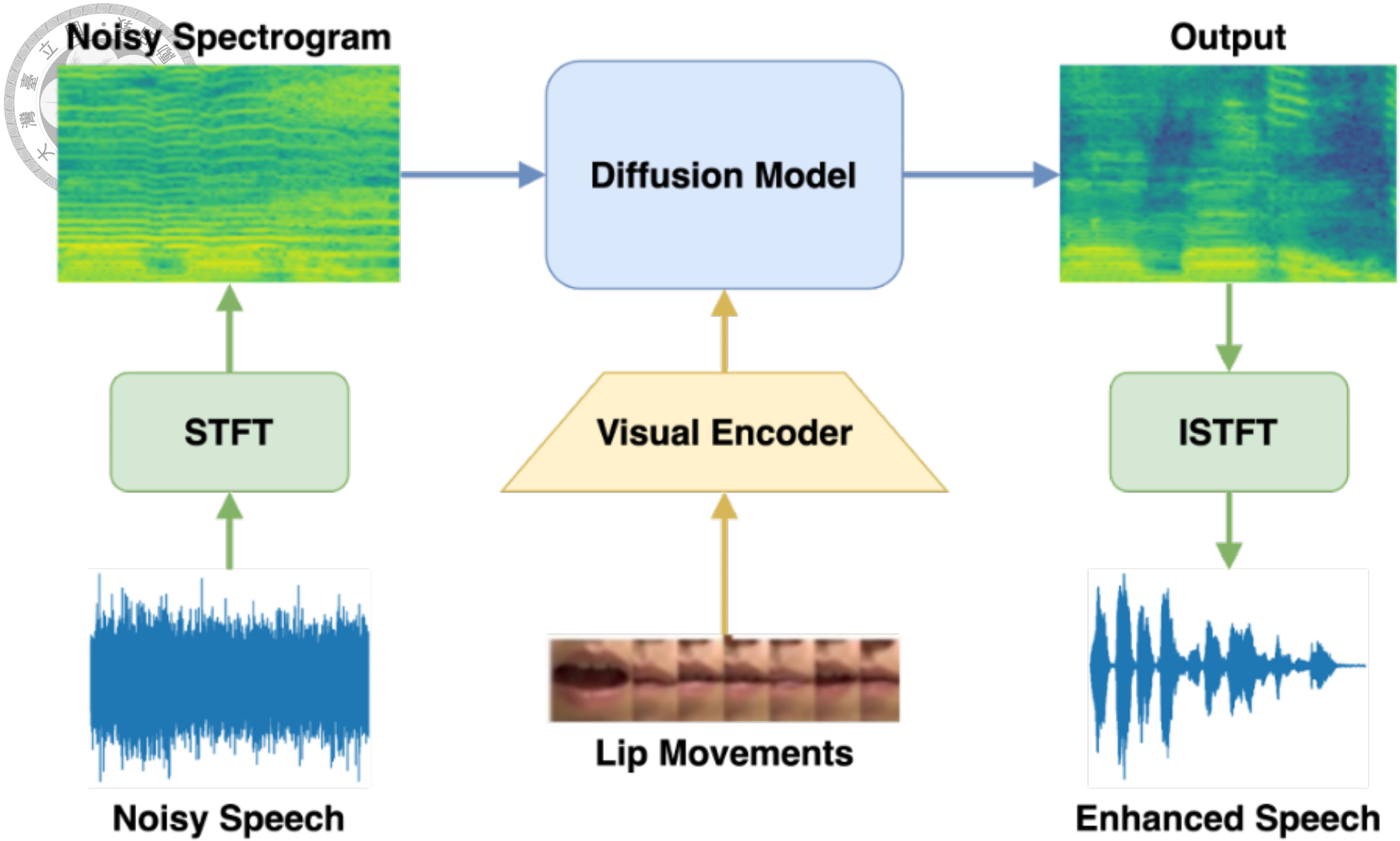


Figure 3.1: Overview of our proposed DAVSE system

diffusion steps, progressively refining the spectrogram towards a cleaner version.

Concurrently, visual information is extracted from the video frames. Specifically, lip movement data provides visual cues that can help resolve ambiguities and fill in missing details in the audio signal. This visual information is processed by a pre-trained visual encoder [29], which extracts meaningful and high-dimensional embedded features from the speaker's lip region of interest (ROI). Mathematically, given the input video sequence $V = \{v_1, v_2, \ldots, v_T\}$, the visual encoder E_v extracts visual features as:

$$F_v = E_v(V) \tag{3.2}$$

The visual encoder's goal is to capture the temporal dynamics and spatial details of the mouth movements, which are then used to enhance the accuracy of the speech signal recovery.

Cross-attention modules integrate the extracted visual features into the audio enhancement pipeline. Let F_a represent the audio features extracted from the noisy spectrogram by an audio encoder. The cross-attention mechanism is then defined as follows:

$$Q = M_Q(F_a), \quad K = M_K(F_v), \quad V = M_V(F_v) \tag{3.3}$$

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.4}$$

Where d_k is the dimensionality of the keys, and the softmax operation is used to obtain the attention weights, ensuring that the relevant visual features are appropriately

doi:10.6342/NTU202502501

weighted for each audio feature.

Following the denoising process, the enhanced spectrogram is converted back into a waveform using the inverse Short-Time Fourier Transform (ISTFT). This transformation allows for a high-quality audio output in the time domain, reconstructing a waveform that has improved intelligibility and preserves natural speech characteristics. The entire system, therefore, integrates a multi-modal approach combining the strengths of diffusion models and visual features, thereby substantially improving speech enhancement compared to audio-only methods.

Algorithm 1 Diffusion-based Audio-Visual Speech Enhancement (DAVSE)

Require: Noisy audio signal $y(t)$, lip movements $V = \{v_1, v_2, \ldots, v_T\}$
Ensure: Enhanced audio signal $y'(t)$
Convert $y(t)$ to spectrogram using STFT:
    $F_a \leftarrow \text{STFT}(y(t))$
Extract visual features:
    $F_v \leftarrow \text{VisualEncoder}(V)$ ▷ VisualEncoder includes 3D Conv, ResNet-18, TCN
Estimate noise using Score-based Diffusion Model:
    $\epsilon_{\text{est}} \leftarrow \text{SGM}(F_a, T, F_v)$ ▷ T is the reverse timestep
Compute enhanced spectrogram:
    $Y_{\text{est}} \leftarrow F_a - \epsilon_{\text{est}}$
Reconstruct audio via ISTFT:
    $y_{\text{est}}(t) \leftarrow \text{ISTFT}(Y_{\text{est}})$
return $y_{\text{est}}(t)$

To provide a better understanding of the step-by-step operation of our system, we present the pseudocode for the proposed DAVSE in Algorithm 1:

The proposed DAVSE system's reliance on audio and visual cues makes it particularly robust in scenarios where speech is heavily corrupted by noise. By iteratively refining the noisy speech spectrogram while being guided by the embedded lip movement information, our system demonstrates an effective means of achieving state-of-the-art performance in audio-visual speech enhancement.

doi:10.6342/NTU202502501

3.1 Visual Encoder

The visual encoder in the Diffusion-based Audio-Visual Speech Enhancement (DAVSE) system is designed to extract meaningful lip movement features that are subsequently used to aid in audio denoising. Our approach uses a pre-trained visual encoder [29] that effectively captures the speaker's mouth region's temporal dynamics and spatial details. The encoder architecture consists of three sequential parts: 3D Convolutional Layers, ResNet-18, and a Temporal Convolution Network (TCN) [6], as shown in Figure 3.2. In the following, we explain the role of each component and its significance in the overall pipeline.

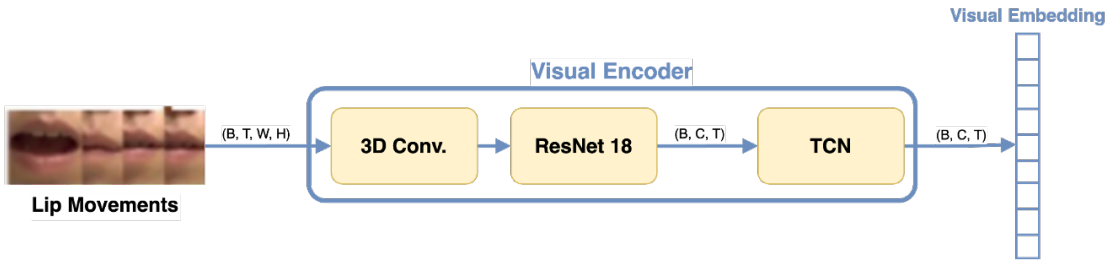


Figure 3.2: The structure of the visual encoder

3.1.1 3D Convolutional Layers

The initial component of the visual encoder comprises 3D Convolutional Layers. These layers are critical for capturing the spatio-temporal characteristics of lip movements. Given the input of frames from the lip region, the 3D convolution operation can extract both spatial information (the shape and appearance of the lips) and temporal changes (the movement over time). 3D convolution layers are essential because lip movements are not just a static shape; they involve dynamic changes over time, reflecting the articulation of different speech sounds.

The output of the 3D convolution layers retains the temporal consistency of the input

doi:10.6342/NTU202502501

sequence while embedding detailed spatial features of the mouth's movement. This facilitates the downstream components by providing enriched spatio-temporal representations, crucial for accurate speech enhancement, particularly in scenarios where the audio signal is heavily corrupted by noise.

3.1.2 ResNet-18 Model

Following the 3D convolutional layers, the visual encoder utilizes ResNet-18 to extract higher-level features from the spatiotemporal representations. ResNet-18 is a well-established convolutional neural network architecture that incorporates residual connections to ease the training of deeper networks. The inclusion of ResNet-18 in the visual encoder serves a dual purpose: to compress the spatial information into higher-dimensional feature embeddings while preserving crucial spatial dependencies and to provide a more abstract representation of the lip movement patterns.

Residual connections within ResNet-18 are vital for mitigating the vanishing gradient problem, allowing deeper networks to converge faster and perform better. In the context of visual speech enhancement, ResNet-18 effectively transforms the rich, low-level visual features extracted by the 3D convolution layers into a compact, abstract representation that is more suitable for guiding the audio denoising process.

3.1.3 Temporal Convolution Network (TCN)

The final component of the visual encoder is a Temporal Convolution Network (TCN), which focuses on modeling the temporal dependencies of the encoded lip features. Unlike traditional RNNs or LSTMs typically used for sequence modeling, TCNs have several

doi:10.6342/NTU202502501

advantages in terms of parallelization and the ability to model long-range dependencies. TCNs use causal convolution layers with residual connections, ensuring that predictions at each time step are only influenced by past inputs. This is particularly important for maintaining the sequential integrity of temporal features.

In the DAVSE system, the TCN processes the output of ResNet-18 to capture the temporal relationships across the entire sequence of lip movements. The encoder then produces an embedding that captures how each frame of the lip movement relates to the others in time, thus capturing speech dynamics effectively. These visual embeddings are crucial for aligning the visual information with the audio spectrogram, ultimately helping the diffusion model to enhance noisy audio by utilizing both current and historical lip motion context.

3.1.4 Visual Embedding Output

The output of the visual encoder is a temporal embedding that represents the dynamic changes in lip movements over time. This visual embedding is subsequently used as a condition within the cross-attention mechanism of the diffusion model. By encoding the temporal dynamics of the lip region, the visual encoder ensures that the subsequent speech enhancement stages have access to contextually rich, speaker-specific visual features that directly correspond to spoken content. The embedding helps the DAVSE model disambiguate and reconstruct noisy speech signals, especially in challenging scenarios involving severe acoustic interference.

doi:10.6342/NTU202502501

3.2 Cross-Attention Mechanism

The cross-attention mechanism in the Diffusion-based Audio-Visual Speech Enhancement (DAVSE) system is designed to optimally fuse the audio and visual modalities, allowing the model to generate an enhanced audio signal guided by complementary visual cues. The structure of the cross-attention module is illustrated in Figure 3.3. In this section, we explain how the features from both modalities are fused and outline the advantages of this approach over traditional feature fusion methods.

3.2.1 Cross-Attention Structure

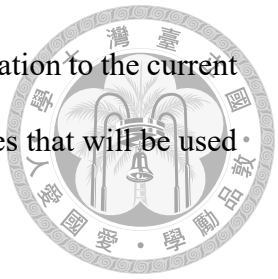
The cross-attention module takes two primary inputs: the *audio features* derived from the noisy spectrogram and the *visual embeddings* extracted from the visual encoder. The attention mechanism works by forming three distinct components: **Query**, **Key**, and **Value**, each of which plays a critical role in aligning the relevant information across modalities.

- **Query Generation (M_Q)**: The noisy audio spectrogram features serve as the input to a transformation layer (M_Q), which is responsible for generating the **Query**. This component allows the audio input to direct the focus of the attention mechanism, determining which audio segments need guidance from the visual modality for denoising.

- **Key and Value Generation (M_K and M_V)**: The visual embedding generated by the visual encoder is passed through two different transformation layers (M_K and M_V), producing the **Key** and **Value** respectively. The **Key** provides a set of reference

doi:10.6342/NTU202502501

features that can be used to assess the relevance of visual information to the current audio frame, while the **Value** represents the actual visual features that will be used to augment the audio.



- **Attention Block**: The **attention block** computes a compatibility score between the **Query** (audio features) and **Key** (visual features) to determine which visual features are most relevant to the current segment of the audio signal. These scores are then used to create a weighted combination of the **Value** vectors (visual features), which effectively introduces visual cues at appropriate moments to guide the denoising of the noisy audio signal. This fusion process ensures that the enhanced audio incorporates critical visual information that enhances clarity and intelligibility.

3.2.2 Benefits of Cross-Attention over Other Fusion Methods

The cross-attention mechanism offers several advantages over more traditional feature fusion methods such as concatenation, addition, or early/late fusion techniques:

- **Dynamic Feature Weighting**: Unlike simple concatenation or addition of features, cross-attention allows the model to dynamically adjust the importance of visual features for each audio segment. The flexibility is particularly beneficial when visual information has varying levels of importance depending on the context—e.g., when the lips are clearly visible or when different phonetic elements require more or less visual guidance.

- **Efficient Feature Fusion Without High Computational Overhead**: Cross-attention effectively fuses two different modalities—audio and visual—without incurring significant additional computational costs. Unlike other advanced fusion methods re-

doi:10.6342/NTU202502501

quiring complex operations or multiple neural network branches for various features, the cross-attention mechanism leverages a transformer-like architecture that efficiently aligns and integrates the features through learned attention weights. This approach ensures the computational burden remains manageable while achieving a high-quality audio and visual information fusion. Furthermore, using learned query, key, and value transformations allows for targeted feature selection, minimizing redundant operations and avoiding unnecessary computational complexity compared to techniques that perform brute-force fusion.

Cross-attention ensures an efficient and contextually relevant combination of modalities by dynamically weighting and fusing the visual cues into the audio spectrogram. This results in better-enhanced speech quality without compromising computational efficiency.

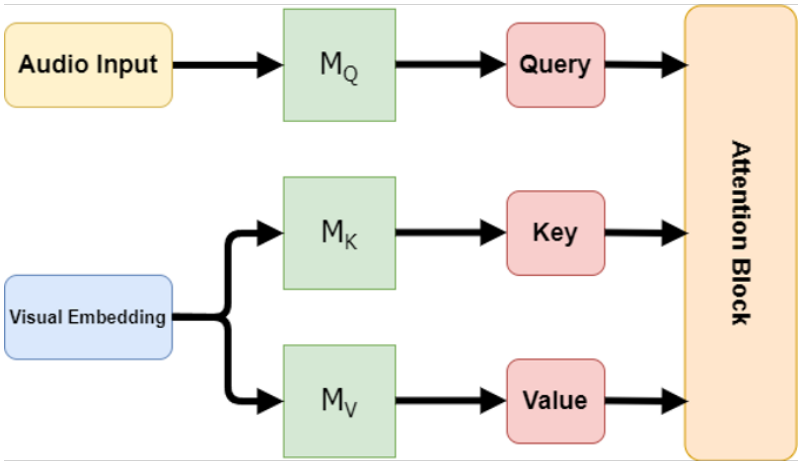


Figure 3.3: Cross-Attention Module for Audio-Visual Feature Fusion.

3.2.3 Diffusion Model Details

Our core model for denoising is based on a modified version of the Noise Conditional Score Network (NCSN++) [38], which serves as the backbone of our Spectrogram Generative Modeling Speech Enhancement (SGMSE) [32]. The diffusion model is tailored to

doi:10.6342/NTU202502501

iteratively remove noise from the spectrogram, progressively refining it through multiple reverse diffusion steps; the structure is shown in 3.4.

The diffusion process is formulated as a Markov chain that gradually transforms a noisy spectrogram into an enhanced version by minimizing the differences between predicted and actual noise across several iterations. Each step of the reverse process applies a score-based estimation technique, utilizing learned noise approximations to guide the denoising trajectory toward a clean spectrogram. Our modifications to the SGMSE structure include the integration of cross-attention modules, which effectively introduce the visual modality into each denoising step. This modification ensures that the model can condition its predictions on previously denoised states and relevant visual features at each denoising iteration, leading to better temporal coherence and intelligibility.

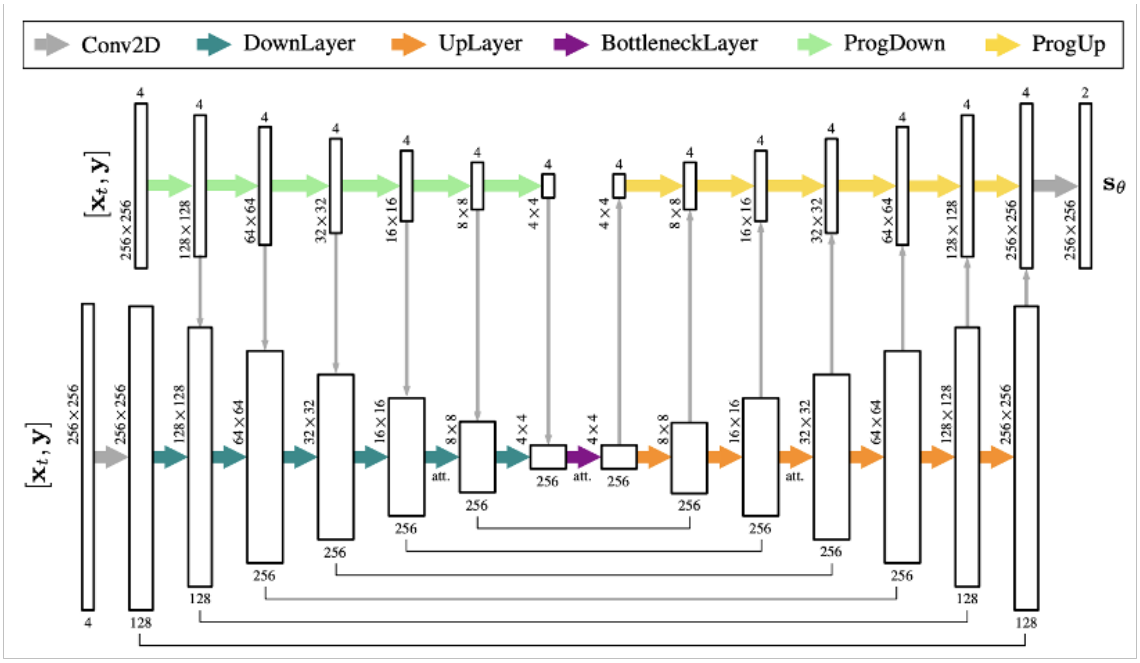


Figure 3.4: The architecture of NCSN++.
[38]

In addition, we made another critical modification to the architecture by **removing the attention layers in the upsampling and downsampling blocks**. We found no need to introduce visual features at different resolutions during the upsampling and downsam-

doi:10.6342/NTU202502501

pling. This change helped streamline the model by focusing the computational resources on key stages where the visual features provide the most significant enhancement effect while reducing the model's overall complexity and computational overhead.

The input spectrogram is generated via the short-time Fourier transform (STFT), allowing for effective frequency-domain processing, which the diffusion model can directly manipulate. After denoising, the spectrogram is transformed back to the time domain using inverse STFT (ISTFT) to obtain the final enhanced speech signal.

3.3 Drift-Supervised Loss

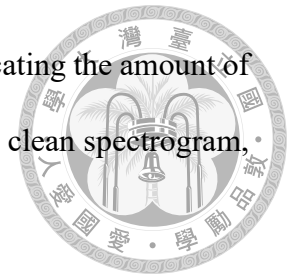
To further improve the performance of the Diffusion-based Audio-Visual Speech Enhancement (DAVSE) system, we introduce a novel drift-supervised loss inspired by the generative-supervised learning loss described in [5]. Specifically, we incorporate visual information as an additional input to guide the diffusion process. Moreover, instead of using L_2 Norm (MSE) as the generative-supervised learning loss did, we use L_1 Norm (MAE) in our implementation.

In [5], a weighted generative-supervised learning loss was proposed to enhance the quality of the generated audio during the diffusion process. We adapt this concept by extending the loss formulation to consider the visual features derived from lip movements, providing additional context for the speech enhancement process. This visual information is invaluable for improving model performance in challenging, noisy environments.

The loss function for our proposed DAVSE model consists of a weighted combination of the original score-based model loss and a newly introduced drift-supervised loss, which supervises the model during the reverse diffusion process. The drift-supervised

doi:10.6342/NTU202502501

loss minimizes the discrepancy between the predicted gradients (indicating the amount of noise to be removed) and the actual gradients derived from the target clean spectrogram, effectively guiding the model at each denoising iteration.



Formally, let:

- $\epsilon_\theta(F_t, t, F_v)$ be the estimated noise by the model, conditioned on both audio and visual features.
- ϵ represents the actual noise component.

The drift-supervised loss is expressed as follows:

$$\mathcal{L}_{\text{drift}} = \mathbb{E}_{F_t, F_v}\left[||\epsilon - \epsilon_\theta(F_t, t, F_v)||\right] \tag{3.5}$$

In summary, the objective loss function for our DAVSE model is a weighted combination of the score-matching loss and the drift-supervised loss:

$$\mathcal{L}_{\text{DAVSE}} = \alpha\mathcal{L}_{\text{Score}} + (1-\alpha)\mathcal{L}_{\text{Drift}} \tag{3.6}$$

Where α represents the predefined weight that controls the contribution of the score-matching loss and drift-supervised loss.

The drift-supervised loss ensures that the model learns an accurate path from the noisy spectrogram toward the clean spectrogram by introducing visual features as an additional conditioning factor. This approach is particularly effective in stabilizing the training process and guiding the reverse diffusion steps, allowing the model to precisely enhance the

doi:10.6342/NTU202502501

speech signal even in challenging noise scenarios.

By integrating visual guidance through cross-attention and leveraging a weighted combination of generative and drift supervision, our DAVSE system achieves state-of-the-art performance in audio-visual speech enhancement, providing more transparent and more natural-sounding speech even under adverse acoustic conditions.

doi:10.6342/NTU202502501

Chapter 4 Experiments

This chapter presents the experimental setup, preprocessing steps, datasets used, evaluation metrics, and results obtained for our proposed Diffusion-based Audio-Visual Speech Enhancement (DAVSE) system. We detail the implementation settings, discuss the baseline models used for comparison, and analyze the performance of DAVSE across different metrics.

4.1 Experimental Setup

4.1.1 Datasets

We evaluated the DAVSE system using two datasets that provide both audio and visual speech recordings:

- **TMSV Dataset**: A Chinese dataset consisting of 18 speakers, with 320 video samples of 2 seconds each.
- **3rd AVSE Challenge Dataset extracted from LRS3**: An English dataset comprising 37,823 speakers, each with a 2 to 10 seconds video clip. This dataset allows us to test our model on diverse speech scenarios.

doi:10.6342/NTU202502501

For both datasets, noisy versions of the clean audio were created by mixing clean speech with different types of noise corpus at signal-to-noise ratios (SNRs) ranging from -10 dB to 10 dB.

4.1.2 Audio Preprocessing

The audio preprocessing pipeline, which is shown in Figure 4.1, consists of several steps aimed at preparing the noisy speech for enhancement:

- **Resample**: All audio signals are resampled to a uniform sampling rate of 16 kHz to maintain consistency across datasets.
- **Pad/Crop**: Each audio sample is padded or cropped to a pre-defined target length to ensure uniform input length for batch processing.
- **Short-Time Fourier Transform (STFT)**: The audio is converted into a spectrogram representation using the STFT with a window size of 510, resulting in 256 frequency bins, a hop length of 160, and a periodic Hann window.
- **Amplitude Transformation**: An amplitude transformation is applied to STFT coefficients to account for the commonly heavy-tailed distribution of STFT speech amplitudes.

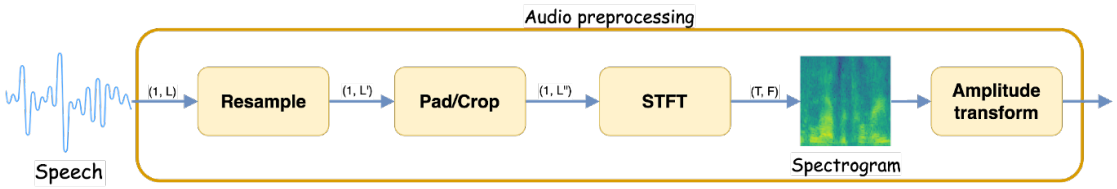


Figure 4.1: Illustration of the audio preprocessing pipeline

doi:10.6342/NTU202502501

4.1.3 Video Preprocessing

The video preprocessing pipeline focuses on extracting meaningful features from the visual input:

- **Frame Extraction**: Video frames are extracted at 25 frames per second (fps) to align with the temporal resolution of the audio.
- **Face Detection**: The Mediapipe face tracker is used to detect faces in the frames.
- **Landmark Identification and Face Alignment**: The Face Alignment Network (FAN) is used to identify facial landmarks. Size and rotation differences are corrected by aligning faces to the canonical coordinates of the mean face of the training set.
- **Mouth Region Extraction**: A bounding box of size 96×96 pixels is used to crop the mouth regions of interest (ROI).
- **Normalization**: Each frame is normalized by subtracting the mean and dividing by the standard deviation of the training set.

4.1.4 Two-stage Training

When dealing with large-scale datasets like LRS3, the initialization of the diffusion process can significantly impact the quality of the enhanced speech. To address this issue, we adopt a two-stage training process proposed by Lemercier *et al*. [25], where an estimate provided by a predictive model serves as a guide for further diffusion, as illustrated in Figure 4.2. This approach has been proven to improve the model's performance by providing

doi:10.6342/NTU202502501

a more accurate initial estimate, resulting in higher quality speech enhancement [24].

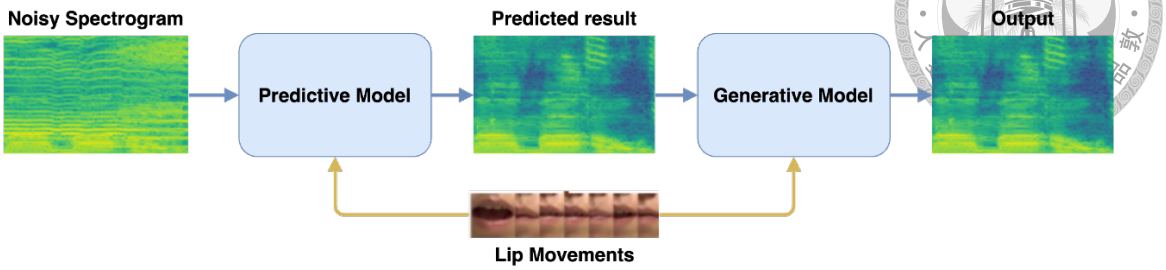


Figure 4.2: Illustration of two-stage training

4.2 Baselines

To establish the performance of our DAVSE system, we compared it against several baseline models:

- **Audio-only Baseline**: A model using the same diffusion approach but without leveraging any visual information, providing a benchmark for the contribution of visual data.
- **ILAVSE** [11]: An improved, lightweight audio-visual speech enhancement system fusing audio and visual features for speech enhancement.
- **AVCVAE** [34]: Using variational methods to model the uncertainty in both audio and visual data.
- **SSL AVSE** [22]: An audio-visual model using a self-supervised learning model (SSL) to enhance speech using both audio and visual features.

The baselines were chosen to provide a range of comparisons, from purely audio-based methods to state-of-the-art audio-visual fusion techniques.

doi:10.6342/NTU202502501

4.3 Evaluation Metrics

We used the following metrics to evaluate the quality and intelligibility of the enhanced speech:

- **Perceptual Evaluation of Speech Quality (PESQ)** [33]: Evaluates the quality of the processed speech compared to the clean speech, providing a score ranging from 0.5 to 4.5.
- **Short-Time Objective Intelligibility (STOI)** [39]: Measures the intelligibility of the enhanced speech with values ranging from 0 to 1, where higher scores indicate better intelligibility.
- **Scale-Invariant Signal-to-Distortion Ratio (SI-SDR)** [23]: Assesses the quality of speech by comparing the enhanced speech to a clean reference, focusing on minimizing the impact of gain variations. It is widely used for evaluating the performance of speech separation and enhancement models. Higher values indicate better performance.

4.4 Results

4.4.1 Quantitative Results

Table 4.1 summarizes the quantitative results of our DAVSE system compared to the baseline models across different noise levels. Our results demonstrate that incorporating visual information significantly enhances speech quality and intelligibility, particularly under challenging low SNR conditions.

doi:10.6342/NTU202502501

Method	PESQ	STOI
Noisy	1.19	0.6
AVCVAE	1.34	0.63
SSL AVSE	1.4	0.68
ILAVSE	1.41	0.64
Audio-only Baseline	1.53	0.69
DAVSE	**1.79**	**0.73**

Table 4.1: Comparison of audio-visual speech enhancement models on the TMSV dataset.

Method	PESQ	STOI
Without Visual Information	1.53	0.69
Without Drift-Supervised Loss	1.7	0.69
Without MAE Loss	1.75	0.73
Full Model	**1.79**	**0.73**

Table 4.2: Ablation study results showing the effect of removing visual information and drift-supervised loss on performance.

4.4.2 Ablation Study

We conducted an ablation study to assess the contribution of different components of our DAVSE system, such as the use of visual information and drift-supervised loss. The results are presented in Table 4.2, highlighting that removing visual information led to a significant performance drop across all metrics.

4.4.3 Generalization Ability

To evaluate the generalization capability of DAVSE, we tested it on the 3rd AVSE Challenge dataset extracted from LRS3, which contains a diverse set of speakers and speaking conditions. The results are summarized in Table 4.3. Despite the variability in the dataset, DAVSE maintains superior performance compared to the baseline method, demonstrating its robustness across different speakers and noise conditions.

doi:10.6342/NTU202502501

Method	PESQ	STOI
Noisy	1.46	0.61
Baseline	1.72	0.69
DAVSE	**1.96**	**0.71**

Table 4.3: Speech enhancement results tested on LRS3 dataset.

Loss type	PESQ	STOI
MSE	1.7	0.69
MAE	**1.73**	**0.71**

Table 4.4: Comparison of different loss types on the TMSV dataset.

4.4.4 Visualization

We provide a qualitative assessment through the visualization of spectrograms of noisy, clean, and enhanced speech in Figure 4.3 and Figure 4.4. Our DAVSE system effectively reduces noise and retains important speech characteristics, particularly in low-SNR environments.

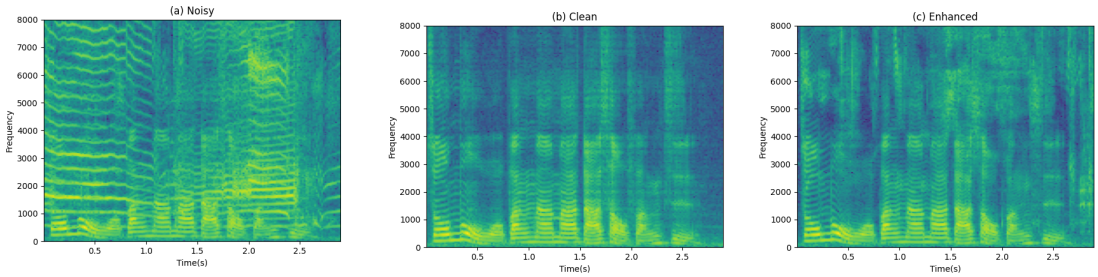


Figure 4.3: Spectrogram comparison for a sample utterance in TMSV: (a) Noisy input, (b) Clean target, (c) Enhanced by DAVSE.

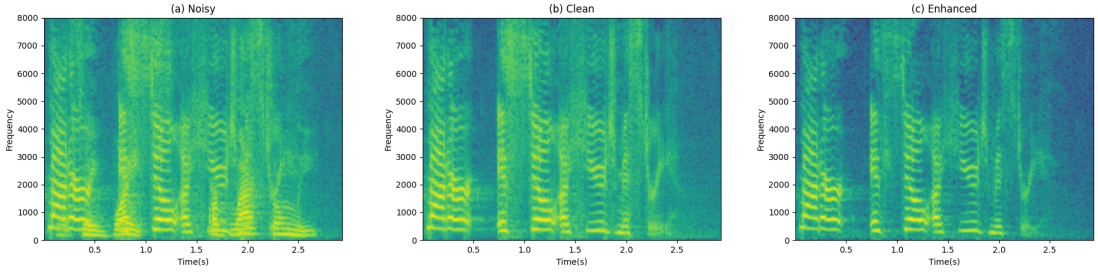


Figure 4.4: Spectrogram comparison for a sample utterance in LRS3: (a) Noisy input, (b) Clean target, (c) Enhanced by DAVSE.

doi:10.6342/NTU202502501

Method	PESQ	STOI
with two-stage training	1.39	0.65
w/o two-stage training	**1.7**	**0.69**

Table 4.5: Comparison of different training processes on TMSV dataset.

Resolution	PESQ	STOI
(16, 8, 4, 0)	1.62	**0.71**
(8, 0)	1.66	0.7
(0)	**1.7**	0.69

Table 4.6: Comparison of adding attention layers at different resolutions during training on TMSV dataset.

4.5 Discussion

The experimental results demonstrate that our DAVSE system consistently outperforms the baseline models in both objective and subjective measures. Including visual information through a dedicated visual stream significantly improves the quality and intelligibility of the enhanced speech, which is evident from the increased scores in both PESQ and STOI.

The ablation studies clearly indicate the contribution of each major component within the DAVSE system. Removing visual information led to a significant drop in performance, highlighting the importance of integrating lip movements to provide context when enhancing speech in noisy environments. The visual modality plays a crucial role, especially in low-SNR conditions where audio cues alone are insufficient to reconstruct clean speech.

Similarly, the drift-supervised loss proves to be essential for the model's effectiveness. By supervising the diffusion process with additional guidance on the drift, the model achieves a more stable and accurate enhancement trajectory. This results in better noise reduction and improved preservation of speech details. Without this drift-supervised loss, the model showed a moderate decline in performance, indicating that the loss effectively

doi:10.6342/NTU202502501

aligns the generated outputs with the clean target.

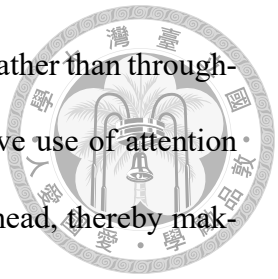
We also explored the impact of different types of loss functions. As shown in Table 4.4, using Mean Absolute Error (MAE) instead of Mean Squared Error (MSE) improved the enhancement performance. The MAE loss helps in reducing the influence of significant outliers that can skew the error measurement, leading to a more robust training process and better overall quality of the enhanced speech.

The generalization study on the LRS3 dataset demonstrates that the DAVSE model can effectively enhance speech across diverse speaking scenarios. Even though LRS3 contains a much wider variety of speakers and environments, the DAVSE model maintains a high level of performance compared to the baseline methods. This ability to generalize well to unseen data underlines the robustness of our approach, making it suitable for real-world applications.

We also compared the effect of using a two-stage training method on the TMSV dataset. Interestingly, the results in Table 4.5 show that the DAVSE model performs better **without** two-stage training in this setting. This supports our hypothesis that two-stage training is more beneficial for large and diverse datasets, such as LRS3, where a good initialization is essential due to high variability across speakers and scenarios. In contrast, smaller and more homogeneous datasets like TMSV appear to be more robust and can benefit from training directly from pure noise initialization. This finding highlights the importance of adapting training strategies to the characteristics of the dataset.

Furthermore, the experiments comparing the affection of adding attention layers at different resolutions (Table 4.6) revealed that attention mechanisms significantly enhance the model's performance when used at specific resolutions. This suggests that the model

doi:10.6342/NTU202502501



benefits from a careful balance of applying attention at specific layers rather than throughout the entire upsampling and downsampling processes. This selective use of attention layers improves the performance and reduces the computational overhead, thereby making the model more efficient.

Finally, the spectrogram visualizations (Figures 4.3 and 4.4) provide a qualitative assessment of DAVSE's capability to reduce noise and retain important speech characteristics effectively. In the noisy spectrogram, speech components are often masked by noise, making them indistinguishable. However, after enhancement by DAVSE, the enhanced spectrogram closely resembles the clean target, with noise components significantly attenuated and the primary speech features clearly visible. The results showed that our model is effective in quantitative metrics and qualitatively produces high-quality, natural-sounding speech.

In conclusion, the integration of visual cues, the drift-supervised learning approach, and the attention mechanisms at specific resolutions contribute significantly to the success of the DAVSE system. These components collectively enhance the system's ability to effectively suppress noise and retain speech quality, even under challenging conditions. The results obtained on both controlled and generalizable datasets validate the efficacy of the proposed model, suggesting its potential applicability for real-world scenarios, such as assistive communication devices, video conferencing, and other noisy environments where clear speech is crucial.

doi:10.6342/NTU202502501

Chapter 5 Conclusion

In this work, we proposed the Diffusion-based Audio-Visual Speech Enhancement (DAVSE) framework, which leverages a diffusion model to enhance speech by conditioning on noise-robust latent visual representations. The DAVSE system is designed to effectively utilize audio and visual cues, significantly improving speech quality and intelligibility, especially in challenging, noisy environments.

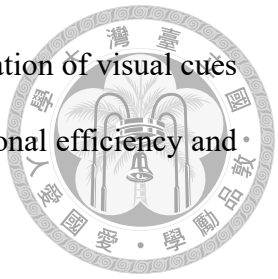
The key contributions of this work are summarized as follows:

- **Proposed the DAVSE Framework**: We introduced a novel framework incorporating diffusion-based modeling for Audio-Visual Speech Enhancement (AVSE). Our approach effectively conditions the diffusion process on noise-robust latent visual representations, ensuring that the visual stream provides meaningful contextual information that enhances the speech signal. This conditioning helps improve the system's robustness, particularly in scenarios where the audio is heavily corrupted by noise.

- **Audio-Visual Feature Fusion via Cross-Attention Mechanism**: The fusion of audio and visual features was achieved using a cross-attention mechanism, allowing the model to dynamically integrate visual information without incurring significant additional computational costs. Unlike traditional fusion techniques, the cross-

doi:10.6342/NTU202502501

attention mechanism ensures efficient and context-aware integration of visual cues into the enhancement process. This approach retains computational efficiency and significantly improves the quality of the enhanced speech.



- **Effectiveness of Drift-Supervised Loss**: We introduced and demonstrated the effectiveness of a novel drift-supervised loss in the AVSE setting. This loss term provides guidance during the diffusion process, helping the model better align the noisy and clean spectrograms, thereby improving the denoising capability of the system. The experimental results showed that the drift-supervised loss is crucial in stabilizing the training process and enhancing the final output quality.

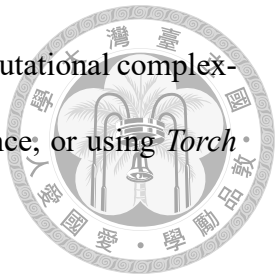
- **Achieved State-of-the-Art Results**: Our DAVSE system achieved state-of-the-art results on the TMSV dataset, outperforming both audio-only and other audio-visual baseline models. The evaluation metrics, including Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI), demonstrated significant improvements, validating the benefits of incorporating both visual cues and advanced diffusion techniques into the speech enhancement pipeline.

5.1 Limitations

While the proposed DAVSE system achieves significant improvements, there are a few limitations that need to be addressed:

- **Inference Latency**: The current inference latency of the DAVSE system is relatively long, averaging approximately 2 seconds per sample. This latency makes the model impractical for real-time applications. To address this, future work can

doi:10.6342/NTU202502501



explore optimizations such as *latent diffusion* to reduce the computational complexity, *bfloat16* to decrease precision while maintaining performance, or using *Torch compile* features to optimize model execution.

- **Residual Noise at the End of Speech**: The system sometimes fails to eliminate noise towards the end of speech segments completely. This issue may be due to the limitations in current audio preprocessing techniques. Future investigations will focus on trying different preprocessing methods that could lead to more consistent denoising across entire speech samples.

5.2 Future Work

To further improve upon the DAVSE system, several directions for future research are suggested:

- **Combining with Text Information**: One promising future direction is incorporating textual information into the enhancement process. By integrating Automatic Speech Recognition (ASR) output or using a text-conditioned speech synthesis model, the system could gain additional context that allows it to further improve the quality and intelligibility of enhanced speech, particularly in challenging acoustic environments.

- **Optimization for Real-Time Applications**: Given the current inference latency, future work will explore ways to make DAVSE suitable for real-time applications. This could involve using more lightweight architectures or employing model quantization techniques.

doi:10.6342/NTU202502501

5.3 Conclusion

Overall, the proposed DAVSE framework pushes the boundaries of current AVSE methods by utilizing advanced diffusion modeling, cross-attention for efficient feature fusion, and a drift-supervised learning approach to guide the enhancement process. The experimental results on both the TMSV and LRS3 datasets have demonstrated the robustness and generalizability of our model, showcasing its potential for real-world applications.

The ability of DAVSE to leverage visual information effectively while keeping computational costs manageable makes it suitable for deployment in various practical scenarios, such as hearing aids, mobile devices, and video conferencing systems, where maintaining high speech quality in noisy environments is crucial. Future work will continue to address the current limitations and explore the integration of additional data modalities, with the ultimate goal of creating a more versatile and efficient AVSE solution.

doi:10.6342/NTU202502501

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech enhancement. arXiv preprint arXiv:1809.02108, 2018.

[2] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[3] T. Afouras, J. S. Chung, and A. Zisserman. Conversation transcription using neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6204–6208. IEEE, 2018.

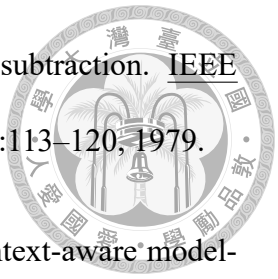
[4] B. D. Anderson. Reverse-time diffusion equation models. Stochastic Processes and their Applications, 12(3):313–326, 1982.

[5] J.-E. Ayilo, M. Sadeghi, and R. Serizel. Diffusion-based speech enhancement with a weighted generative-supervised learning loss, 2023.

[6] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018.

[7] J. Benesty, M. M. Sondhi, and Y. Huang. Speech enhancement. Springer, 2005.

doi:10.6342/NTU202502501



[8] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(2):113–120, 1979.

[9] J. Chen and Y. Luo. Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. IEEE Journal of Selected Topics in Signal Processing, 14(3):423–433, 2020.

[10] N. Chen, Y. Zhang, H. Liu, and P. Smaragdis. WaveGrad 2: Iterative refinement for speech denoising. In ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing, 2022.

[11] S.-Y. Chuang, H.-M. Wang, and Y. Tsao. Improved lite audio-visual speech enhancement, 2022.

[12] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, 33(2):443–445, 1985.

[13] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Transactions on Speech and Audio Processing, 3(1):3–20, 1995.

[14] A. Ephrat, I. Mosseri, R. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, M. Rubinstein, and S. Peleg. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In ACM Transactions on Graphics (TOG), volume 37, pages 1–11, 2018.

[15] A. Gabbay, B. Shillingford, Y. M. Assael, T. Paine, and D. Warde-Farley. Visual speech enhancement. In Interspeech 2018, 2018.

doi:10.6342/NTU202502501

[16] L. Girin, X. Li, M. Bousse, and X. Alameda-Pineda. Davis: a corpus for audiovisual speech processing in presence of noise. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(12):2045–2058, 2019.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (NeurIPS 2020), 2020.

[19] J. Hou, S. Chen, X. Liu, L. Wang, and H. Yin. AVSEGAN: Audio-visual speech enhancement using generative adversarial networks. IEEE Transactions on Multimedia, 2021.

[20] Z. Hou, T. Wang, and L. Ding. Audio-visual speech enhancement using deep neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 290–294. IEEE, 2018.

[21] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In International Conference on Learning Representations (ICLR 2021), 2021.

[22] R. L. Lai, J.-C. Hou, M. Gogate, K. Dashtipour, A. Hussain, and Y. Tsao. Audio-visual speech enhancement using self-supervised learning to improve speech intelligibility in cochlear implant simulations, 2023.

[23] J. Le Roux, F. Weninger, J. R. Hershey, and B. Schuller. Sdr – half-baked or well done? In ICASSP, page 471–475, 2019.

doi:10.6342/NTU202502501

[24] S. Lee, C. Jung, Y. Jang, J. Kim, and J. S. Chung. Seeing through the conversation: Audio-visual speech separation based on diffusion model. In Proc. ICASSP, pages 12632–12636, 2024.

[25] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann. Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.

[26] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. Prentice-Hall, 1979.

[27] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao. Conditional diffusion probabilistic model for speech enhancement. In Proc. ICASSP, pages 7402–7406, 2022.

[28] Y. Luo and N. Mesgarani. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. In IEEE/ACM Transactions on Audio, Speech, and Language Processing, volume 27, pages 1256–1266. IEEE, 2019.

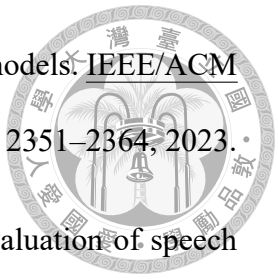
[29] P. Ma, Y. Wang, S. Petridis, J. Shen, and M. Pantic. Training strategies for improved lip-reading. In Proc. ICASSP, 2022.

[30] D. Michelsanti, Z.-H. Tan, J. Jensen, and X.-L. Zhang. An overview of audio-visual speech enhancement. IEEE Journal of Selected Topics in Signal Processing, 2021.

[31] S. Pascual, A. Bonafonte, and J. Serrà. Segan: Speech enhancement generative adversarial network. In Interspeech, pages 3642–3646, 2017.

[32] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann. Speech en-

doi:10.6342/NTU202502501



hancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, pages 2351–2364, 2023.

[33] A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), volume 2, pages 749–752 vol.2, 2001.

[34] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud. Audio-visual speech enhancement using conditional variational auto-encoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2020.

[35] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction, 2022.

[36] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265, 2015.

[37] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020.

[38] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations, 2021.

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing, 19(7):2125–2136, 2011.

doi:10.6342/NTU202502501

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30, 2017.

[41] D. Wang and J. Chen. Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10):1702–1726, 2018.

[42] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Processing Letters, 21(1):65–68, 2014.

doi:10.6342/NTU202502501

Appendix A — Temporal Convolution Networks

Temporal Convolutional Networks (TCNs) are convolutional models specifically designed for sequence modeling tasks. Unlike recurrent architectures, TCNs rely entirely on convolution operations to capture temporal dependencies, offering advantages such as parallel training, stable gradients, and flexible receptive fields. Based on the framework proposed by Bai et al. [6], a TCN is composed of three key components: causal convolutions, dilated convolutions, and residual connections.

A.1 Causal Convolutions

To preserve the temporal order of the input sequence, TCNs utilize causal convolutions, where the output at time step t depends only on inputs from time steps $\leq t$. This is implemented by zero-padding on the left side of the input, thereby preventing the use of future information. Thanks to this causality, TCNs are particularly well-suited for autoregressive modeling and real-time prediction tasks.

doi:10.6342/NTU202502501

A.2 Dilated Convolutions

To efficiently model long-term dependencies, TCNs adopt dilated convolutions. In a dilated convolution, fixed intervals are inserted between the elements of the convolution kernel, enabling an expanded receptive field without significantly increasing the number of parameters or layers. The dilation factor typically increases exponentially across layers (e.g., $d = 1, 2, 4, ...$), allowing the network to cover longer temporal ranges with fewer layers.

A.3 Residual Blocks

Each layer in a TCN is implemented as a residual block, consisting of two dilated causal convolution layers, each followed by weight normalization, ReLU activation, and dropout. A residual connection adds the input of the block directly to its output, and a 1×1 convolution is applied when necessary to match the dimensionality. This design facilitates gradient flow and supports the effective training of deep networks.

doi:10.6342/NTU202502501

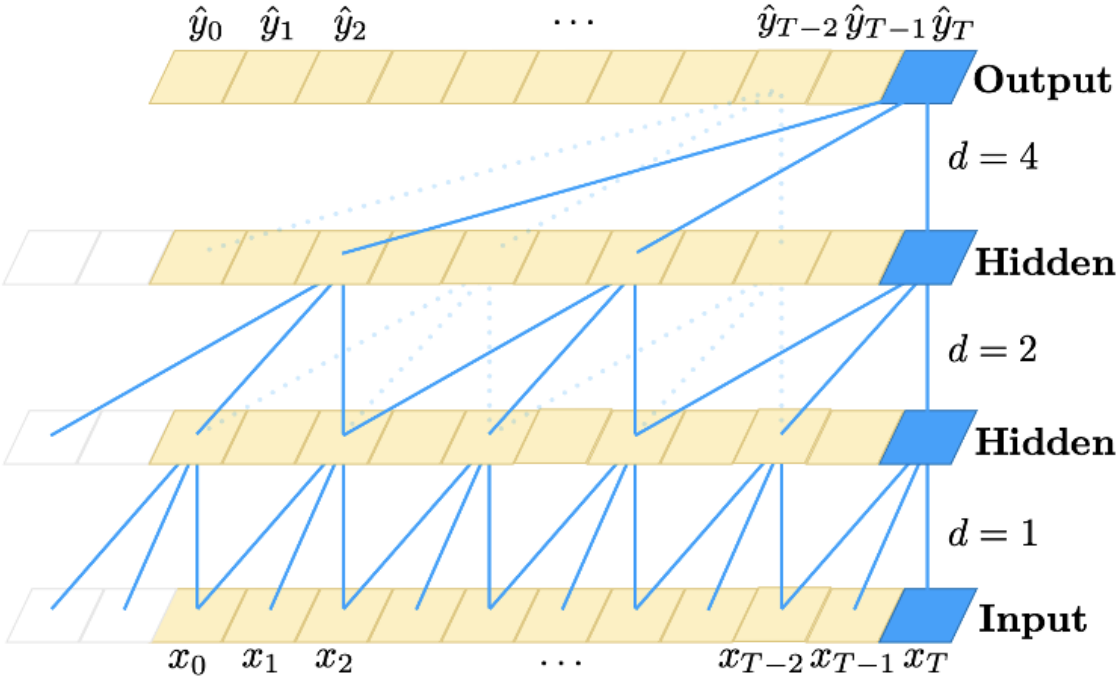


Figure A.1: The structure of a temporal convolution network (TCN) [6]

doi:10.6342/NTU202502501

doi:10.6342/NTU202502501

Appendix B — Derivation of the Stochastic Differential Equation (SDE)

This work follows the continuous-time diffusion framework proposed by Song et al. [38], in which generative modeling is defined via stochastic differential equations (SDEs). This formulation provides a unifying and flexible mathematical foundation for score-based generative models.

B.1 Forward SDE

Let $\mathbf{x}(t) \in \mathbb{R}^d$ denote a data sample at time $t \in [0, T]$. The forward process gradually perturbs the data distribution $p_0(\mathbf{x})$ into a known prior distribution (typically standard Gaussian) via an Itô SDE of the form:

$$d\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), t)\, dt + g(t)\, d\mathbf{w}(t), \tag{B.1}$$

where $\mathbf{f}(\mathbf{x}, t)$ is the drift term, $g(t)$ is the diffusion coefficient, and $\mathbf{w}(t)$ denotes a standard Wiener process. This forward SDE defines a continuous family of marginal distributions $\{p_t(\mathbf{x})\}_{t\in[0,T]}$ that evolve from data to noise.

doi:10.6342/NTU202502501

B.2 Reverse-Time SDE

To generate new samples, we consider the reverse-time stochastic process that transforms noise back into data. According to Anderson [4], under mild regularity conditions, the reverse-time SDE is given by:

$$d\mathbf{x}(t) = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] dt + g(t)\, d\bar{\mathbf{w}}(t), \tag{B.2}$$

where $\bar{\mathbf{w}}(t)$ is a standard Wiener process run backward in time, and $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is the score function of the marginal distribution at time t.

B.3 Score-Based Approximation

Since the true score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is intractable in practice, we approximate it using a neural network $s_\theta(\mathbf{x}, t)$ trained via score matching. Substituting the learned score into the reverse SDE yields the following sampling dynamics:

$$d\mathbf{x}(t) = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2 s_\theta(\mathbf{x}, t)\right] dt + g(t)\, d\bar{\mathbf{w}}(t). \tag{B.3}$$

This learned reverse process can be simulated using numerical solvers (e.g., Euler–Maruyama) to generate data by integrating from $t = T$ to $t = 0$.

This unified SDE framework allows flexible control over the noise injection and sampling trajectory, enabling stable and expressive generative modeling across various domains.

doi:10.6342/NTU202502501