SEEING THROUGH THE CONVERSATION: AUDIO-VISUAL SPEECH SEPARATION BASED ON DIFFUSION MODEL

Suyeon Lee, Chaeyoung Jung*, Youngjoon Jang, Jaehun Kim, Joon Son Chung*

Korea Advanced Institute of Science and Technology, South Korea

ABSTRACT

The objective of this work is to extract the target speaker's voice from a mixture of voices using visual cues. Existing works on audio-visual speech separation have demonstrated their performance with promising intelligibility, but maintaining naturalness remains challenging. To address this issue, we propose AVDiffuSS, an audio-visual speech separation model based on a diffusion mechanism known for its capability to generate natural samples. We also propose a cross-attention-based feature fusion mechanism for an effective fusion of the two modalities for diffusion. This mechanism is specifically tailored for the speech domain to integrate the phonetic information from audio-visual correspondence in speech generation. In this way, the fusion process maintains the high temporal resolution of the features, without excessive computational requirements. We demonstrate that the proposed framework achieves state-of-the-art results on two benchmarks, including VoxCeleb2 and LRS3, producing speech with notably better naturalness. Project page with demo: https://mm.kaist.ac.kr/projects/avdiffuss/

***Index Terms*—** diffusion, stochastic differential equation, audio-visual, speech separation

1. INTRODUCTION

While significant advancements in audio-only speech recognition and separation techniques have been witnessed recently, challenges remain in understanding a speech from an individual amidst overlapping sounds. In real-world situations, conversations are often intertwined with other voices or disturbed by a cacophony of noises. Elimination of such disturbances is crucial in settings like meetings, where one has to focus on the speech of a single individual. Humans excel at guiding their attention to a sound source of interest in such environments, naturally de-emphasizing other sounds. The importance of visual modality in humans' understanding of spoken communications is emphasized in instances where auditory cues contradict visual cues from the speaker's face, leading to frequent misinterpretation of speech sounds [1].

Audio-Visual Speech Separation (AVSS) aims to emulate this human capacity, distinguishing each voice from a collective soundscape using visual information. Beyond enhancing the auditory intelligibility for listeners, this technique can also serve as a preprocessing step for various speech-related tasks, including cascaded speech recognition [2, 3] and speaker diarization [4]. Consequently, there have been significant advances in audio-visual speech separation, driven by the accessibility of multi-modal datasets and high-performance computing. Early works [2, 5] have proposed to combine visual and audio features to distinguish the target speaker's speech in complex and noisy environments. A noteworthy finding in their research is that leveraging the visual modality effectively addresses the label permutation problem, which arises from the challenge of assigning a proper ground truth to the predicted output during training. Recently, VisualVoice [6] utilizes both lip motions and facial attributes (e.g. gender, age, and nationality) as conditions to specify the target speaker. Thus, it is reported that leveraging lip movements is effective for aligning auditory and visual information to extract phonetic information, and incorporating facial attributes aids in distinguishing target speakers using their identity cues.

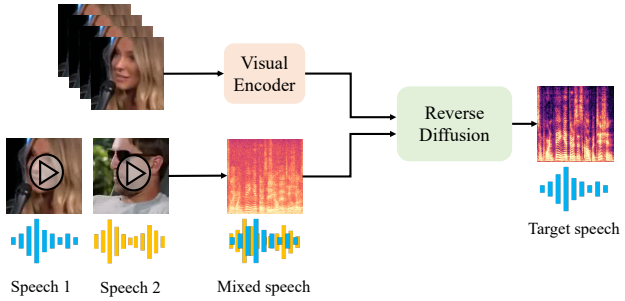

Fig. 1. A pipeline for audio-visual speech separation based on a diffusion model. Mixed speech and target speaker's face crop images are used as inputs for the model, which then extracts the target speaker's speech through a reverse diffusion process.

With the advancements in deep learning, there have been successful applications of generative models in the AVSS field. Generative AVSS models [7, 8] produce realistic samples by learning the mapping from latent space to clean speech distribution. Although these approaches have demonstrated successful performance, they face difficulties in generating diverse samples and exact data estimation, frequently producing speech with undesirable artifacts. This indicates the need for generating samples that sound more natural to humans. In response to this, we take advantage of the generation capabilities of the diffusion-based generative model. The diffusion model proposed and developed in earlier works [9, 10, 11] is known for its potential in generating diverse and natural samples across various domains [12, 13, 14, 15], including the audio-only speech separation [16, 17].

In this work, we propose a diffusion-based AVSS model called `AVDiffuSS` that reconstructs natural and intelligible utterances. We fuse audio and visual modalities in the generation process of speech to incorporate extra information from the video. We mitigate the requirement of pre-training of the encoder and decoder in the widely-used feature fusion strategy [12]. Our method supports a frequency-domain compression of audio features without any extra processing step, enabling end-to-end training.

Our contributions consist of the following: (1) To the best of our knowledge, we are the first to introduce an audio-visual speech sep-

* These authors contributed equally.

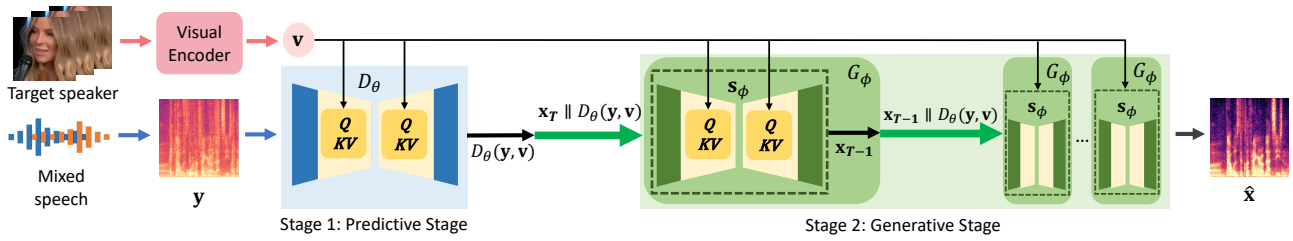

Fig. 2. Model architecture of AVDiffuSS. Face-cropped images of the target speaker are fed to the visual encoder to obtain the visual embedding $\mathbf{v}$. A mixture of two speech signals is transformed into a spectrogram $\mathbf{y}$ by STFT, and it goes into two stages: 1) a predictive stage, and 2) a generative stage. Each stage consists of U-net architecture with cross-attention layers. For the input of the generative stage, the output of first stage $D_\theta(\mathbf{y},\mathbf{v})$ is concatenated with the $\mathbf{x_T}$, which is sampled from $\mathcal{N}(D_\theta(\mathbf{y},\mathbf{v}),\, \sigma^2\mathbf{I})$. In the next stage, reverse diffusion is repeated for N steps at inference. Note that $\parallel$ denotes concatenation.

aration based on a diffusion model, capable of reconstructing both natural and intelligible speech. (2) With the help of the proposed compressing strategy, we successfully mitigate the excessive computational overheads and make our model suitable in the speech domain. (3) With various experiments, we demonstrate that the proposed method attains state-of-the-art results on two widely used benchmark datasets.

2. METHOD

As illustrated in Fig. 2, our framework comprises two main stages: the predictive stage and the generative stage. In the predictive stage, the model initially estimates the speech of the target speaker using visual semantics $\mathbf{v}$ extracted by the visual encoder. The output of the predictive stage, denoted as $D_\theta(\mathbf{y},\mathbf{v})$, is then fed into the generative stage, which employs a diffusion-based model. In this stage, the initial prediction is further enhanced through an iterative denoising process. Note that both stages improve audio-visual alignments by utilizing a task-specific cross-attention module, resulting in the generation of more natural samples.

2.1. Visual Encoder

The visual modality plays two pivotal roles in audio-visual speech separation: (1) synchronizing speech with lip movements to capture phonetic details, and (2) identifying the target speaker based on facial attributes, such as gender, age, and nationality. Taking inspiration from a study in active speaker detection [18], a visual encoder capable of both preserving temporal dynamics and incorporating visual cues can be leveraged to achieve the aforementioned objectives. We adopt the encoder architecture from [18], which comprises a series of ResNet18 layers and a temporal convolutional network from [19]. On top of those modules, a 1D convolution layer is attached to reduce the channel dimension. The encoder, as a result, outputs frame-level spatio-temporal features.

2.2. Encoder-Decoder Free Conditioning by Cross-Attention

To effectively separate the desired speech by exploiting the visual modality, it is essential to maintain the temporal characteristics of both auditory and visual features throughout the fusion process. Based on this, we focus on the cross-attention mechanism, which enables the model to learn the correspondence between sequential information from the two different modalities. Since cross-attention calculates correlations between different modalities by multiplication, it requires heavy computational costs. In response to this, we propose a feature fusion method using cross-attention, eliminating the need for a complex feature compression process involving encoder-decoder architecture.

The proposed feature fusion method is conducted in both predictive and generative stages. In each stage, we aim to acquire the correspondence between visual and audio embeddings. As we adopt the U-net architecture as the backbone of both stages, audio embedding can be represented as $\mathbf{e}_{a,i} \in \mathbb{R}^{C_i \times T_i \times F_i}$. Here, C_i, T_i, and F_i denote the number of channels, frame lengths, and frequency lengths, respectively, in the i-th U-net layer. By applying frequency-axis pooling to the audio features, we obtain the pooled audio feature denoted as $\bar{\mathbf{e}}_{a,i} \in \mathbb{R}^{C_i \times T_i}$, which is used as the query, while the visual feature is employed as the key and value. The output of the cross-attention module is repeated F_i times to recover the original shape of the input before averaging across frequencies. Through this process, our model denoises undesired speech while enhancing the voice of the target speaker.

2.3. Audio-Visual Speech Separation with Diffusion

Predictive stage. In the first stage, a predictive model D_θ predicts the separated speech in one pass. The aforementioned visual encoder and the cross-attention mechanisms are utilized in this stage to use both modalities in the separation process. The initial prediction $D_\theta(\mathbf{y},\mathbf{v})$ serves as a conditioning factor for the second stage.

Generative stage. The diffusion process in the generative stage consists of two procedures, forward process and reverse process. During the forward process $\{\mathbf{x}_\tau\}_{\tau=0}^{T}$ indexed by a continuous time variable τ, a data $\mathbf{x}$ undergoes a gradual perturbation by adding Gaussian noise from a clean data $\mathbf{x}_0$ to a noisy data $\mathbf{x}_T$. A diffusion model [10, 20] is designed to reverse this process, ultimately generating a clean data point $\mathbf{x}_0 \sim p_0$ from a noisy prior $\mathbf{x}_T \sim p_T$. In the context of score-based generative models, a score model, denoted as $\mathbf{s}_\phi$, is trained to estimate $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, which corresponds to the logarithm of the data density function's gradient with respect to the data $\mathbf{x}$. The forward process is mathematically modeled using a Stochastic Differential Equation (SDE) [20], defined as follows:

$$\mathrm{d}\mathbf{x}_\tau = f(\mathbf{x}_\tau,\tau)\mathrm{d}\tau + g(\tau)\mathrm{d}\mathbf{w}. \tag{1}$$

Here, $\mathbf{w}$ represents the standard Wiener process, f acts as the drift coefficient of $\mathbf{x}_\tau$, and g serves as the diffusion coefficient controlling the magnitude of additional Gaussian noise at each step. Our model utilizes SDE from the class of Ornstein-Uhlenbeck SDEs [21], with the drift coefficient f defined as $f(\mathbf{x}_\tau,\tau) := \gamma(D_\theta(\mathbf{y},\mathbf{v}) - \mathbf{x}_\tau)$, where γ indicates a stiffness parameter. This SDE has been applied in previous works for speech enhancement [15, 13]. As τ progresses from 0 to T in the forward

Method	A-V	Diff	PESQ	ESTOI	SI-SDR
DiffSep [16]		✓	2.2070	0.6080	4.4070
VisualVoice [6]	✓		1.9586	0.7696	9.5757
AVDiffuSS (Ours)	✓	✓	**2.5906**	**0.8152**	**12.2701**

Table 1. Speech separation results on the VoxCeleb2 dataset. For all metrics, higher is better. A-V refers to the audio-visual model, and Diff refers to the diffusion-based model.

process in Eq. (1), $\mathbf{x}_\tau$ approaches $D_\theta(\mathbf{y}, \mathbf{v})$ with accumulated Gaussian noise. The reverse-diffusion process in our model is guided by the initial prediction $D_\theta(\mathbf{y}, \mathbf{v})$ which is trained to be similar to ground truth $\mathbf{x}$. In the reverse process, the model is trained to solve the corresponding reverse-time SDE, which is expressed as:

$$\mathrm{d}\mathbf{x}_\tau = \left[f(\mathbf{x}_\tau, \tau) - g(\tau)^2 \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)\right] \mathrm{d}\tau + g(\tau)\mathrm{d}\overline{\mathbf{w}}, \quad (2)$$

where $\overline{\mathbf{w}}$ denotes a standard Wiener process representing the reverse time flow from T to 0, with $\mathrm{d}\tau$ signifying an infinitesimal negative time step. Through the reverse process, a natural and intelligible speech $\hat{\mathbf{x}}$ is generated from a noisy prior $\mathbf{x}_T$.

Training objective. For the joint training of predictive model D_θ and generative stage G_ϕ, we adopt a multi-task learning strategy as described in [13]. The following equations indicated in Eq. (3)-(5) outline the overall training process. The predictor D_θ is trained to directly separate the desired speech from the noisy speech $\mathbf{y}$ with the aid of $\mathbf{v}$. The loss function for D_θ is denoted in L_{pred}, which is an L2 loss computed between the initial prediction $D_\theta(\mathbf{y}, \mathbf{v})$ and the ground-truth $\mathbf{x}$. The loss function for the second stage L_{diff} is determined for the given timestep τ uniformly sampled from $[0, T]$. The score model $\mathbf{s}_\phi$ is trained with a denoising score matching objective [22] introduced in [20], regarding the noise scale σ_τ for time step τ. With weight values λ_1 and λ_2 for these two objectives, respectively, the total loss L_{total} for balanced training is as follows:

$$L_{pred} = \mathbb{E}\left[\|\mathbf{x} - D_\theta(\mathbf{y}, \mathbf{v})\|_2^2\right], \quad (3)$$

$$L_{diff} = \mathbb{E}\left[\|\mathbf{s}_\phi(\mathbf{x}_\tau, \mathbf{y}, \mathbf{v}, \tau) + \frac{\mathbf{x}_\tau - \mathbf{x}}{\sigma_\tau}\|_2^2\right], \quad (4)$$

$$L_{total} = \lambda_1 * L_{pred} + \lambda_2 * L_{diff}. \quad (5)$$

It is important to note that G_ϕ encompasses each step of the reverse-diffusion stage, involving both the estimation of a score by $\mathbf{s}_\phi(\mathbf{x}_\tau, \mathbf{y}, \mathbf{v}, \tau)$ and the sampling procedure to obtain $\mathbf{x}_{\tau-1}$.

Inference procedure. In the predictive stage, the model $D_\theta(\cdot)$ generates an initial prediction of the target speech using $\mathbf{y}$ and $\mathbf{v}$. For the second stage, we can set the number of reverse-diffusion steps N, which controls the step size between each diffusion timesteps τ, thereby affecting the quality of the generated output.

The output of the first stage $D_\theta(\mathbf{y}, \mathbf{v})$ is employed to sample $\mathbf{x}_T$ from the distribution $\mathcal{N}(\mu_T, \sigma_T^2\mathbf{I})$ for the second stage, where μ_T is set as $D_\theta(\mathbf{y}, \mathbf{v})$. The input for the score model $\mathbf{s}_\phi$ is constructed by a concatenation of $D_\theta(\mathbf{y}, \mathbf{v})$ and $\mathbf{x}_T$. By solving the Eq. (2), $\mathbf{x}_{T-1}$ is obtained. $\mathbf{x}_{T-1}$ is then fed into G_ϕ alongside $D_\theta(\mathbf{y}, \mathbf{v})$, and this process is repeated for N steps. Through this reverse-diffusion process, the target audio prediction $\hat{\mathbf{x}}$ can be regenerated from $D_\theta(\mathbf{y}, \mathbf{v})$ with improved naturalness and clarity.

3. EXPERIMENTS

We evaluate our model quantitatively using three established speech evaluation metrics, which are Perceptual Evaluation of Speech Quality (PESQ) [23], Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [24], Extended Short-Time Objective Intelligibility (ESTOI) [25], and qualitatively through Mean Opinion Score (MOS).

Method	A-V	Diff	PESQ	ESTOI	SI-SDR
DiffSep [16]		✓	2.0569	0.6810	5.3140
VisualVoice [6]	✓		1.7719	0.7412	7.2274
AVDiffuSS (Ours)	✓	✓	**2.8106**	**0.8856**	**14.1707**

Table 2. Speech separation results on the LRS3 dataset. Results are obtained by the models trained on the VoxCeleb2 dataset.

3.1. Experimental Setup

Datasets. VoxCeleb2 [26] dataset is a widely-used dataset for audio-visual tasks comprising more than 1 million utterances extracted from YouTube videos. This dataset consists of 5,994 identities in the training set, and 118 identities in the test set. Our model is trained on the VoxCeleb2 train set, and 10 utterances in the test dataset are randomly chosen for validation. LRS3 [27] dataset is another popular dataset for audio-visual speech recognition and speech separation. The dataset is made up of 4,004 videos for training and validation, and 412 videos for test sets, which are from TED and TEDx videos.

Implementation details. We utilize NCSN++M [13] for the U-net and modify the cross-attention mechanism on U-net from [12]. We follow the details in [13] for the diffusion process and set the number of reverse-diffusion steps N to 30. The input for the visual encoder is a sequence of face-cropped grayscale images resized to 112×112. Our model is updated with Adam optimizer [28] with an exponential moving average of network parameters with a decay of 0.999 [29], and the learning rate is initialized to 10^{-4}. The weight values λ_1 and λ_2 for L_{pred} and L_{diff} are both set to 0.5. We use 4 RTX A5000 GPUs for training with an effective batch size of 16. We train our network for 30 epochs, which takes approximately 24 days.

Comparison methods. We compare our method with two publicly-available state-of-the-art speech separation models. DiffSep[1] [16] is an audio-only diffusion-based speech separation model extended from SGMSE+ [15], and VisualVoice[2] [6] is an audio-visual speech separation model. We train DiffSep on the VoxCeleb2 dataset from scratch for 20 epochs for pair comparisons, as this results in approximately the same number of iterations as reported in [16]. We also utilize an official pre-trained VisualVoice model and generate a test set, following [6]. Note that every model is trained on the VoxCeleb2 train set and tested on the first 2 seconds of the samples in the test sets of VoxCeleb2 and LRS3 datasets.

3.2. Experimental Results

Quantitative results. To validate the effectiveness of our methods, we show the experimental results on VoxCeleb2 and LRS3 test sets in Table 1 and Table 2, respectively. The importance of the visual modality for accurate separation of the target speech is highlighted in the ESTOI and SI-SDR results of DiffSep. Moreover, our model shows a higher PESQ score than VisualVoice, which indicates that our model generates natural-sounding speech due to the reverse-diffusion stage. To further simulate one-shot speech separation scenarios, we evaluate every model trained with VoxCeleb2 on

[1]https://github.com/fakufaku/diffusion-separation
[2]https://github.com/facebookresearch/VisualVoice

Method	A-V	Diff	MOS
DiffSep [16]		✓	2.24 ± 0.11
VisualVoice [6]	✓		2.98 ± 0.10
AVDiffuSS (Ours)	✓	✓	$\mathbf{4.44 \pm 0.07}$

Table 3. MOS comparison with 95% confidence interval. A group of 17 participants rated 20 lists of audio randomly selected from the results of each model on the VoxCeleb2 dataset.

Method	PESQ	ESTOI	SI-SDR
DiffSep [16]	1.8926	0.5116	0.4439
VisualVoice [6]	1.7675	0.7334	7.4973
AVDiffuSS (Ours)	**2.2387**	**0.7672**	**9.2628**

Table 4. Speech separation results tested on the bottom 30% samples sorted by SI-SDR results of our model on VoxCeleb2 dataset.

the LRS3 test set. The results in Table 2 show the robustness of our model on a cross-dataset evaluation.

Qualitative results. We conduct a subjective evaluation using MOS to measure how much the outputs sound natural to the human ear. A group of 17 participants are asked to assess 20 pairs of separated outputs on a scale of 1 to 5. Every sample is normalized to eliminate the amplitude bias in outputs of each model and the orders of the models are randomly assigned for every pair. Criteria for the evaluation are: (1) audio quality relative to the corresponding ground-truth, and (2) degree of separation. Evaluating the degree of separation is impossible without knowing the ground truth samples when separating the voices of two people with similar tones, because the outcome may sound natural even when the other speaker's speech is included. Therefore, the ground truth samples are provided to the participants as standards for assessing separation capability. As shown in Table 3, the MOS of our model is significantly higher compared to previous works. These results demonstrate the ability of our approach to generate samples that sound clear and natural to human hearing, not to mention their intelligibility.

3.3. Discussions

Experimental results in difficult cases. In real-world scenarios such as a conversation between two speakers with similar timbre, it is difficult to accurately distinguish the speech of each speaker. Thus we show the results from the hardest samples to prove the robustness of our diffusion-based audio-visual approach. Sorted by the SI-SDR result of our model, we choose the bottom 30% samples to demonstrate the performance of each model in harsh conditions, which is potentially disadvantageous to our method. By taking advantage of both diffusion-based approaches and audio-visual ones, our model shows reliable performance even under unfavorable conditions as shown in Table 4. The ESTOI and SI-SDR results of our model and VisualVoice demonstrate the ability of audio-visual models to isolate the whole intelligible speech in harsh cases due to utilizing synchronization cues and facial characteristics. In contrast, DiffSep cannot identify the target speaker accurately due to the lack of visual information, resulting in especially lower scores on SI-SDR.

Spectrograms of the separated outputs from each model are shown in Fig. 3, including ground truth for comparison. A pair of speech signals are randomly selected from the lowest 30% results, and the mixture of speech is fed to each model to evaluate the three models. Boxes and circles with identical colors in each row represent regions that should be the same as the spectrogram of the clean

Resolutions	PESQ	ESTOI	SI-SDR
32	2.2322	0.7177	7.8316
32, 64	2.3687	0.7708	10.1944
32, 64, 128 (Ours)	**2.4984**	**0.7959**	**11.2712**

Table 5. Ablation results on the feature resolutions to which the cross-attention modules are applied. Each model is evaluated on the VoxCeleb2 dataset after 15 epochs of training.

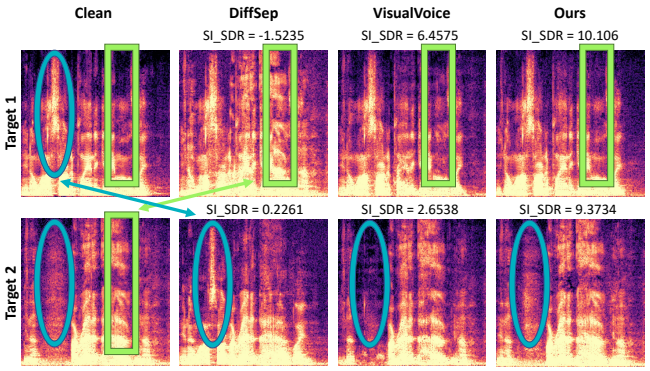

Fig. 3. Spectrogram comparison of the outputs of DiffSep [16], VisualVoice [6] and our model on the lowest 30% random sample.

sample. Arrows colored in green and blue denote that the non-target speaker's speech is included in the output of the audio-only diffusion model, but not in our results. In VisualVoice, the details of the original speech are ignored, and the speech is over-denoised as shown in the spectrograms. This visualization demonstrates the ability of our model to generate realistic details, not to mention the accurate capturing of the spoken contents.

Feature resolution ablations for cross-attention. The U-net layers' feature resolution in our model starts at 256 and is halved four times during the downsampling path, reaching a minimum of 32, and then upsampled to its original size. Among the eight layers, cross-attention modules are applied on up to six layers with the three smallest resolutions. Ablation results in Table 5 show the impact of different feature resolution settings where cross-attention modules are applied. Adding more cross-attention layers leads to a modest performance improvement, indicating the benefits of incorporating audio-visual fusion for speech separation. Yet, we avoid adding cross-attention to every U-net layer due to memory constraints.

4. CONCLUSION

In this work, we present AVDiffuSS, an audio-visual speech separation framework based on the diffusion model. Our approach exploits visual cues to extract the target speaker's speech accurately, and the diffusion model to produce a highly natural-sounding output. We devise a task-specific feature fusion mechanism for integrating a target speaker's visual information. The proposed model demonstrates state-of-the-art performance for audio-visual speech separation in terms of both naturalness and intelligibility.

5. ACKNOWLEDGEMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT, 2022-0-00989).

6. REFERENCES

[1] Harry McGurk and John MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976. 1

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018. 1

[3] Ke Tan, Yong Xu, Shi-Xiong Zhang, Meng Yu, and Dong Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE Journal of Selected Topics in Signal Processing*, 2020. 1

[4] Abudukelimu Wuerkaixi, Kunda Yan, You Zhang, Zhiyao Duan, and Changshui Zhang, "DyViSE: Dynamic vision-guided speaker embedding for audio-visual speaker diarization," in *International Workshop on Multimedia Signal Processing*. IEEE, 2022. 1

[5] 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," in *Proc. ACM SIGGRAPH*, 2018. 1

[6] Ruohan Gao and Kristen Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *Proc. CVPR*, 2021. 1, 3, 4

[7] Karren Yang, Dejan Marković, Steven Krenn, Vasu Agrawal, and Alexander Richard, "Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis," in *Proc. CVPR*, 2022. 1

[8] Viet-Nhat Nguyen, Mostafa Sadeghi, Elisa Ricci, and Xavier Alameda-Pineda, "Deep variational generative models for audio-visual speech separation," in *IEEE 31st International Workshop on Machine Learning for Signal Processing*, 2021. 1

[9] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. ICML*, 2015. 1

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020. 1, 2

[11] Yang Song and Stefano Ermon, "Generative modeling by estimating gradients of the data distribution," in *NeurIPS*, 2019. 1

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR*, 2022. 1, 3

[13] Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2023. 1, 2, 3

[14] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *Proc. ICML*, 2021. 1

[15] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2023. 1, 2, 3

[16] Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaeuk Byun, Soyeon Choe, and Min-Seok Choi, "Diffusion-based generative speech source separation," in *Proc. ICASSP*, 2023. 1, 3, 4

[17] Bo Chen, Chao Wu, and Wenbin Zhao, "SEPDIFF: Speech separation based on denoising diffusion model," in *Proc. ICASSP*, 2023. 1

[18] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," in *Proc. ACM MM*, 2021. 2

[19] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Lipreading using temporal convolutional networks," in *Proc. ICASSP*, 2020. 2

[20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021. 2, 3

[21] B Oksendal, *Stochastic differential equations: an introduction with applications*, Journal of the American Statistical Association, 2000. 2

[22] Pascal Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011. 3

[23] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001. 3

[24] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *Proc. ICASSP*, 2019. 3

[25] Jesper Jensen and Cees H Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2016. 3

[26] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018. 3

[27] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018. 3

[28] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015. 3

[29] Yang Song and Stefano Ermon, "Improved techniques for training score-based generative models," in *NeurIPS*, 2020. 3