Demonstration of LipDiffuser: Lip-to-Speech Generation with Conditional Diffusion Models Running on a Portable Device

Danilo de Oliveira, Julius Richter, Tal Peer, Timo Gerkmann

Signal Processing, University of Hamburg, Hamburg, Germany

***Abstract*—We present a demonstration of *LipDiffuser*, a diffusion-based lip-to-speech model that generates high-quality speech samples using only silent video and an enrollment utterance, all recorded on-the-spot by users. The model runs locally on a laptop with a minimal setup and short processing times, enabling users to assess the quality of the generated speech, including preservation of speaker traits, and explore the limits of lip-reading.**

1. INTRODUCTION

Conditional temporal data generation involves synthesizing data sequences guided by conditioning, such as generating speech signals from corresponding video frames. In particular, conditional diffusion models have shown strong performance in challenging multimodal tasks by learning dependencies between modalities and producing realistic, synchronized outputs [1]–[3].

The lip-to-speech task is especially relevant for restoring audio in cases where the original speech track is either missing (e.g. due to malfunctioning recording equipment) or severely degraded, such that high-fidelity restoration is difficult. Success in this area requires not only intelligible audio reconstruction from silent video footage but also proper alignment of visual and auditory features, in spite of significantly different sampling rates. Additionally, the generated audio should contain the characteristics of the target speaker [4].

The *LipDiffuser* model [5] leverages a conditional diffusion approach to generate natural-sounding, speaker-consistent speech from lip movement videos, outperforming prior methods on standard benchmarks such as the *LRS3* dataset [6]. However, inference results reported in [5] were achieved using high-end hardware and did not include an analysis of computational efficiency or runtime. In this demo, we show that the *LipDiffuser* model can run on a consumer-grade laptop, confirming that the model can efficiently process videos and generate high-quality speech audio in a reasonably short time, without the need for specialized hardware or cloud resources. Our demo shows that state-of-the-art lip-to-speech synthesis is accessible on portable devices, bringing robust speech restoration technology closer to real-world and on-site applications.

2. LIPDIFFUSER

LipDiffuser converts silent video of a speaking person into natural-sounding speech audio, as shown in Figure 1. The model consists of four main components: a video encoder to extract visual speech features from video frames, a speaker encoder to capture speaker identity, a diffusion-based denoiser model that predicts the target mel-spectrogram, and a neural vocoder which generates the final audio waveform from the predicted mel-spectrogram.

During inference, the denoiser model receives four inputs: a noisy intermediate audio representation $\mathbf{x}_t$ (current state of the diffusion process), the video features $\mathbf{v}$, the speaker embedding $\mathbf{s}$, and the diffusion process time step t. It outputs a mel-spectrogram that captures the speech content aligned with the lip movements and speaker characteristics. The spectrogram is then converted into a waveform audio signal by the vocoder and can be merged with the original video.

2.1. Network Architecture

The denoiser network follows the magnitude-preserving ablated diffusion model (MP-ADM) architecture [7], a diffusion-based neural network optimized to preserve the expected magnitudes of activations and model weights for stable training and high output quality. In order to effectively condition on speech cues from video features, which are time-dependent data sequences, the feature-wise linear modulation (FiLM) mechanism [8] is implemented in a time-dependent and magnitude preserving manner (MP-FiLM) [5]. Using MP-FiLM, lip movement information is injected at multiple points within the diffusion network, ensuring that the generated audio is well-synchronized with the video input.

2.2. Training Objective

LipDiffuser is trained using the denoising score matching objective [9]: the model learns to remove progressively added Gaussian noise from spectrogram data while being conditioned on the corresponding video, speaker embeddings and the time step of the noise scheduler. At each training step, the model predicts the original clean spectrogram from a noisy version, aiming to minimize the difference between its prediction and the true data.

2.3. Inference

In order to generate audio conditioned on lip movements, the lip region-of-interest (ROI) is first cropped from video frames and converted to grayscale, resulting in the input features $\tilde{\mathbf{v}} \in \mathbb{R}^{N \times H \times W}$, where N is the number of frames of height H and width W. Subsequently, video features $\mathbf{v} = E_v(\tilde{\mathbf{v}})$ are extracted using the video encoder E_v. A speaker embedding vector $\mathbf{s} = E_s(\tilde{\mathbf{s}})$ is extracted using the speaker encoder E_s, where $\tilde{\mathbf{s}} \in \mathbb{R}^{n_{\tilde{\mathbf{s}}}}$ is the enrollment data of the target speaker's voice containing $n_{\tilde{\mathbf{s}}}$ samples, e.g. a pre-recorded sentence. With the video features and the speaker embedding processed, the reverse diffusion process is applied using the second-order deterministic sampler from Karras et al. [10] along with the trained denoiser model D_θ. The end result of the reverse diffusion process is a mel-spectrogram $\hat{\mathbf{x}} \in \mathbb{R}^{n_a}$. Finally, the neural vocoder D_a is used to synthesize the time-domain speech signal $\tilde{\mathbf{x}} = D_a(\hat{\mathbf{x}})$, with $\tilde{\mathbf{x}} \in \mathbb{R}^{n_{\tilde{x}}}$ containing $n_{\tilde{x}}$ samples.

3. IMPLEMENTATION DETAILS

3.1. Input and output representation

The model expects an enrollment audio signal sampled at 16 kHz; Resampling is performed if necessary according to the recording device. The speaker embedding vector is produced by the *Wespeaker* encoder D_s [11], which extracts speaker characteristics from an enrollment utterance of arbitrary length, producing a single feature vector of size $n_s = 256$.

Video conditioning inputs are recorded at 25 frames per second. The video frames are processed to obtain the grayscale lip ROI, following the pipeline in [12], [13], which uses the face alignment network from [14] to detect and locate faces. The ROIs of size 88×88 pixels

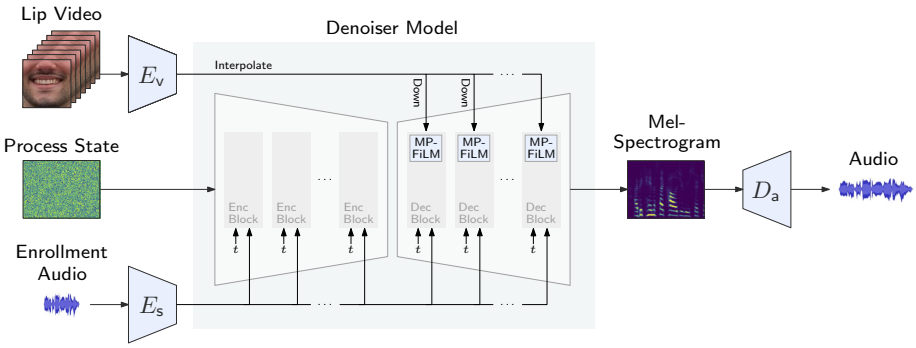

Fig. 1: *LipDiffuser*: The model contains a video encoder E_{v}, a speaker encoder E_{s}, a denoiser model, and a neural vocoder D_{a}. The inputs to the denoiser model are the process state, the current process time t, and the speaker embedding from the speaker encoder and video features extracted fromt the silent video by the video encoder, interpolated to match the audio feature rate, and then downsampled to the feature rate of the decoder's target resolutions. Magnitude-preserving feature-wise linear modulation (MP-FiLM) layers fuse the processed video features within the decoders blocks. After reverse diffusion, the neural vocoder synthesizes an audio waveform signal from the generated mel-spectrogram.

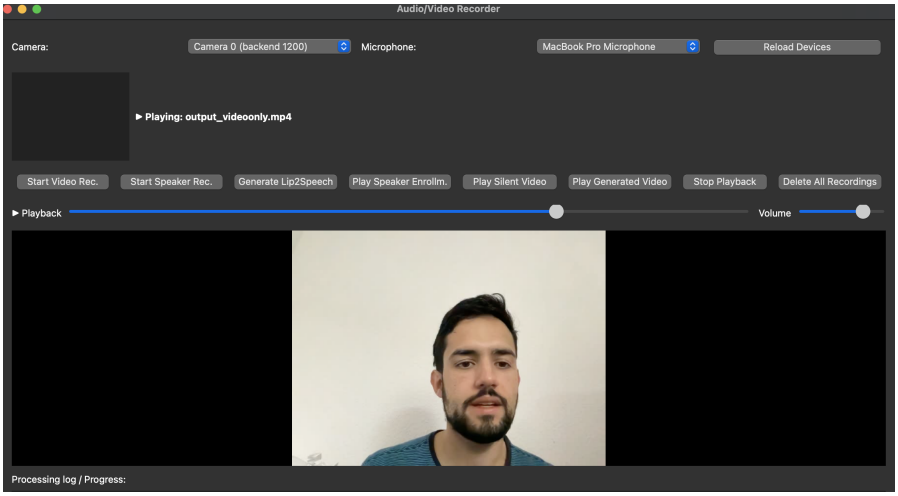

Fig. 2: User interface of the program, showing the buttons for recording silent-video and audio enrollment, generating speech and playing the results.

are cropped and fed into the self-supervised model *BRAVEn* [15], resulting in video features of dimension $n_v = 1024$ per video frame.

The model outputs a mel-spectrogram with an FFT length of 1024, and hop length of 256 samples, thus obtaining a feature rate of 62.5 Hz. The number of mel-frequency bins per frame is $n_a = 80$. The mel-spectrogram is converted into a time-domain audio signal via the neural vocoder D_a, specifically the *HiFi-GAN decoder* [16], pretrained at 16 kHz.

3.2. Model Training

The *Lipdiffuser* used in this demo is trained as described in [5], using the *LRS3* dataset [6], the same data preprocessing pipeline, and the same training strategy.

3.3. Demo Framework

The demo runs locally on a laptop equipped with a GPU. The user interface enables the user to interactively record the video (without sound) and also to record the enrollment utterance. The recording devices can be selected, which can be the microphone and webcam embedded in the laptop itself or external devices. In noisy environments, an optional pretrained speech enhancement model is provided for enhancing the enrollment utterance. After recording the video, the audio generation pipeline is executed. Generation speed can be controlled by the number of steps in the reverse diffusion process, at the expense of audio quality. Finally, the synthesized audio is merged with the original video recording and can be played back to the user directly in the interface.

4. CONCLUSIONS

We present an interactive demo of *LipDiffuser*, a lip-to-speech diffusion model that synthesizes speech from silent video recordings and an enrollment utterance. The demo allows users to assess the quality of generated speech, explore lip-reading through different levels of articulation and observe how speaker characteristics are preserved, depending on enrollment lexical and acoustic content, time duration and noise level in the recording. Furthermore, our demo shows that state-of-the-art lip-to-speech synthesis is possible on portable devices, bringing the technology closer to real-world applications.

REFERENCES

[1] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 8652–8661.

[2] I. Viertola, V. Iashin, and E. Rahtu, "Temporally aligned audio for video with autoregression," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2025, pp. 1–5.

[3] S. Torma and L. Szegletes, "Generative modeling and augmentation of EEG signals using improved diffusion probabilistic models," *Journal of Neural Engineering*, vol. 22, no. 1, p. 016001, 2025.

[4] Y. Yemini, A. Shamsian, L. Bracha, S. Gannot, and E. Fetaya, "LipVoicer: Generating speech from silent videos guided by lip reading," in *Int. Conf. on Learning Representations (ICLR)*, 2024.

[5] D. de Oliveira, J. Richter, T. Peer, and T. Gerkmann, "Lipdiffuser: Lip-to-speech generation with conditional diffusion models," *arXiv preprint arXiv:2505.11391*, 2025.

[6] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[7] T. Karras, M. Aittala, J. Lehtinen, J. Hellsten, T. Aila, and S. Laine, "Analyzing and improving the training dynamics of diffusion models," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 24 174–24 184.

[8] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

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

[10] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 35, pp. 26 565–26 577, 2022.

[11] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2023, pp. 1–5.

[12] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2020, pp. 6319–6323.

[13] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2021, pp. 7613–7617.

[14] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2017.

[15] A. Haliassos, A. Zinonos, R. Mira, S. Petridis, and M. Pantic, "Braven: Improving self-supervised pre-training for visual and auditory speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2024, pp. 11 431–11 435.

[16] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033.