Unsupervised Audio Synthesis from Silent Video Speech

Final Project Report
Maximiliano Medina, Nico Nunez-Sahr, Frederik Stihler

Berkeley EECS
University of California Berkeley

COMPSCI 294 158 (Spring Semester 2024)
Instructors: Pieter Abbeel, Wilson Yan, Kevin Frans, Philipp Wu

1 Introduction

The ability to synthesize accurate speech from silent video footage of people speaking presents a significant technological challenge with widespread practical applications. In scenarios where audio is lost, corrupted, or never originally captured, reconstructing coherent and synchronized speech can provide immense value. The task involves extracting subtle visual cues such as lip movements and facial expressions, and translating these into corresponding speech sounds.

Potential applications of this technology are vast and significant. It can enhance media accessibility for the hearing impaired, ensure clear communication in video calls when audio is compromised, facilitate speech articulation in environments where traditional vocal communication is impractical, and aid forensic analysis by reconstructing audible speech from silent surveillance footage. Each of these applications underscores the transformative impact of this technology across various sectors.

One of the major challenges in synthesizing speech from silent video is creating audio that not only aligns with the speaker's lip movements but also captures their emotional nuances. Achieving realistic speech that conveys emotions rather than sounding robotic hinges largely on the effectiveness of the video embedding. The temporal synchronization is one of the advantages of video-conditioned audio generation over text-based synthesis. Another one is the higher availability of audio-video training data on the internet.

The novelty of this project lies in its innovative use of a diffusion model for audio synthesis from silent video, diverging from the traditional supervised methods like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) that rely on extensive labeled datasets [see Prajwal et al., 2020]. Luo et al. [2023] already introduced the use of a latent diffusion model to generate audio from silent video; however, they do not generate speech and specifically exclude all speech samples from their training data. By adopting unsupervised learning techniques, our approach allows the model to intuitively understand and replicate the complex interactions between visual cues and audio outputs. This not only reduces the need for large annotated datasets, but also improves the model's ability to generalize to diverse and novel video content, representing a significant advance in the field of speech synthesis from visual information.

2 Methodology

2.1 Approach

Our approach integrates an architecture designed to synthesize audio from silent video footage. This system leverages the complementary strengths of two main parts: component 1, a video encoder, and component 2, a diffusion model, each specialized to handle distinct aspects of the audio-visual synthesis process. The model architecture is shown in Figure 1.

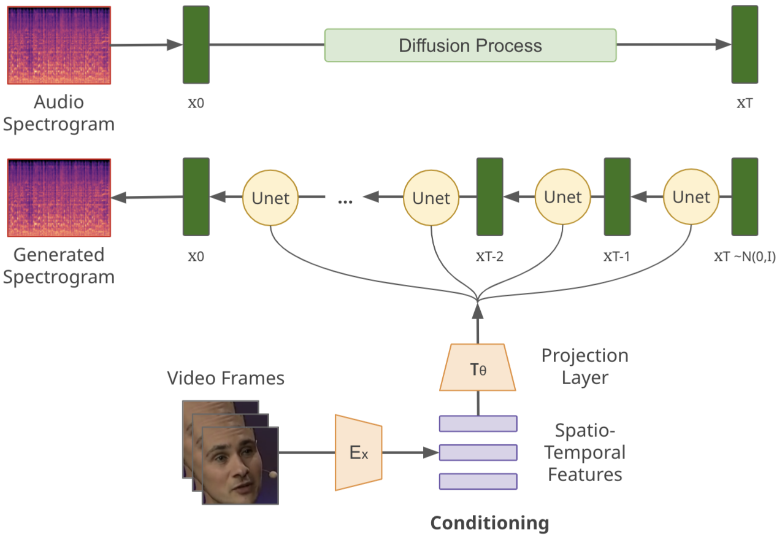

Figure 1: Overview of Architecture for Video Conditioning: Video encoder and Diffusion Model.

2.1.1 Video Encoder

The video encoder is tasked with extracting essential visual features from silent videos that inform the audio generation process. For this purpose, we employ the encoder component of a pretrained Masked Autoencoder (MAE) [Tong et al., 2022]. Videos are presented to the model as 16x16 patches. The pretrained model outputs a sequence of $768-$dimensional vectors that we then use as features for our downstream task. The idea is to utilize these features to enrich our system with a deep, nuanced understanding of visual dynamics, greatly enhancing the synthesis of corresponding audio outputs.

2.1.2 Melspectrogram

Based on the literature [see Prajwal et al., 2020]. Luo et al. [2023], the diffusion model operates not on the 1D soundwaves themselves, but on a 2D representation of the sound called *Melspectrogram*. At a high-level, the Melspectrogram is the rescaled Fourier Transform of the soundwave in the amplitude-time space, obtaining a representation in the log frequency-time space McFee et al. [2015]. This allows us to effectively visualize sound in an image, where the pixel value portrayed as color represents the intensity of the sound in decibels.

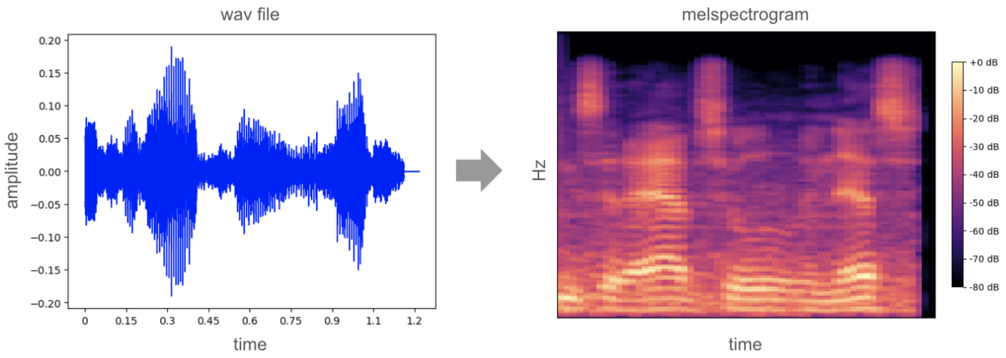

Figure 2: Audio Pre-Processing

2.1.3 Diffusion Model

The diffusion model is central to our approach. It conditions the generation of visual representation of the sound to the video.

Our aim is to train a diffusion model on the Melspectrograms x_0 given a video v_x. More specifically, from v_x we obtain, though a separate pretrained model, an encoded version of the video denoted by E_x. Then, a projection layer τ_θ is used to project up or down E_x into a tensor with a compatible dimension with the diffusion process.

Forward process: In the forward process, Gaussian noise is gradually added to the original Melspectrogram x_0. This operation is performed via a fixed schedule that we denote by $\alpha_1, \cdots, \alpha_T$. T is the total number of timesteps and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The forward process is modeled by function q:

$$\begin{aligned} q(x_t|x_0) &= \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \\ &= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t)I), \end{aligned} \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$. Also

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha}_t x_{t-1}, (1-\alpha_t)I). \tag{2}$$

Loss: We train a U-Net [Ronneberger et al., 2015], denoted by ϵ_θ, to estimate the noise added in each timestep given the encoded video. This architecture has proven robust capabilities in generative tasks. During training we minimize the mean squared error (MSE) between the real and predicted noise as our objective function:

$$L(\theta) = \mathbb{E}_{x_0, t, \epsilon} \|\epsilon - \epsilon_\theta(x_t, t, E_x)\|_2^2. \tag{3}$$

One thing to note is that our implementation combines the encoded video with the timestep by adding both, forcing them to have the same dimension. However, this is an architecture choice that can be evaluated further.

Backward process: The backward process models the reverse of q. During sampling, after the diffusion model is trained, we denoise a sample of Gaussian noise $x_T \sim \mathcal{N}(0, I)$ conditioned on the embedding E_x to obtain a corresponding Melspectrogram. The reverse transition probability can be expressed as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, E_x), \sigma_t^2 I), \tag{4}$$

where the mean and variance are defined as follows:

$$\mu_\theta(x_t, t, E_x) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \epsilon_\theta(x_t, t, E_x) \right), \tag{5}$$

$$\sigma_t^2 = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}(1-\alpha_t). \tag{6}$$

2.2 Data

The data for this project is sourced from the Lip Reading in the Wild (LRW) dataset, available at the Visual Geometry Group from the University of Oxford's website [Chung and Zisserman, 2016]. The LRW dataset is specifically designed for lip-reading tasks and contains up to 1000 utterances of 500 different words, spoken by hundreds of different speakers. Each video in the dataset is 29 frames long, corresponding to 1.16 seconds, with the target word positioned centrally within the clip. The videos are provided in mp4 format.

We obtained access to the LRW dataset through a formal request to the BBC, ensuring compliance with data usage and privacy standards. For the purposes of this project, we focused our research on a subset of the dataset to maintain a manageable scope for model training and testing, while still providing a diverse array of phonetic contexts. We selected 39 classes among the words that start with the letters A through C. The total number of training examples amounts to 41,698.

2.3 Data Pre-processing

The pre-processing of the data involved several steps to prepare the audio and visual components for integration into our model.

Audio: The audio tracks (wav files) were extracted from the corresponding video files. We then converted these audio tracks into Melspectrograms using the librosa package McFee et al. [2015] with a sampling rate of 24,000. To ensure that the Melspectrograms are interpretable by the diffusion model, we pad them at the end to match an appropriate input size (in our case 128x64).

Video: Given the robust capabilities of the pretrained Video Masked Autoencoder (MAE) used for video encoding, no additional pre-processing of the video files was necessary. Because our videos are longer than MAE expects, we apply the MAE to different chunks of the video and concatenate the tensors into a single one.

2.4 Implementation

The model is implemented using Python, leveraging libraries such as PyTorch for model building and training, librosa for audio processing, and ffmpeg for handling video data.

Our project leverages custom classes based on the transformers and diffusers libraries from HuggingFace. We use the UNet2DConditionModel from the diffusers and the pretrained VideoMAE from the transformers library. The UNet is configured with 3 input and 3 output channels suitable for RGB images. Each UNet block contains 2 ResNet layers. The number of output channels for each block varies, with a configuration of (128, 128, 256, 256, 512, 512), allowing for a gradual increase in feature complexity. The model uses a combination of regular ResNet blocks (DownBlock2D and UpBlock2D) and blocks with spatial self-attention (AttnDownBlock2D and AttnUpBlock2D). For inference, we use the DDPM-Scheduler with 1000 sampling steps [Ho et al., 2020].

To aid the training of this complex task, we perform 500 warmup steps and cosine decay of the learning rate until 10^{-4}.

2.5 Evaluation Metrics

For evaluation, we use two primary metrics: Accuracy score and Frobenius Norm.

The accuracy score quantifies the quality of the model's audio generations. We use OpenAI's Whisper-1 [Radford et al., 2022] to convert speech to English text and then check if the ground truth label of the conditioning data is contained in the transcribed text.

It is calculated as:

$$\text{Accuracy} = \frac{\text{Number of labels found in transcriptions}}{\text{Total number of samples}} \tag{7}$$

Although humans can distinguish spoken words in audio files more frequently than the Whisper model, this Speech-to-Text model allowed us to quantify whether our models were generating understandable speech in an automated fashion. It is important to mention that this AI model frequently transcribed spoken words such as 'cancer' into similar-sounding words, like 'councillor', thereby reducing its accuracy versus human listeners. This could be caused by model hallucinations of noisy sounds at the beginning of the generated audio that condition future generated words, making 'councillor' more likely than the true word 'cancer'.

3 Results

Results were gathered from three different models, which conditioned the diffusion process on increasingly higher-dimensional data. The simplest diffusion model conditions data on integer-labeled classes, which are transformed into one-hot encoded tensors, and fed into the trainable projection layer τ_θ to be combined with the 512-dimensional timestep embedding. The second diffusion model conditions the forward and backward processes on the 768-dimensional text embedding of the spoken word. The third model with highest dimensionality conditions the training and sampling processes

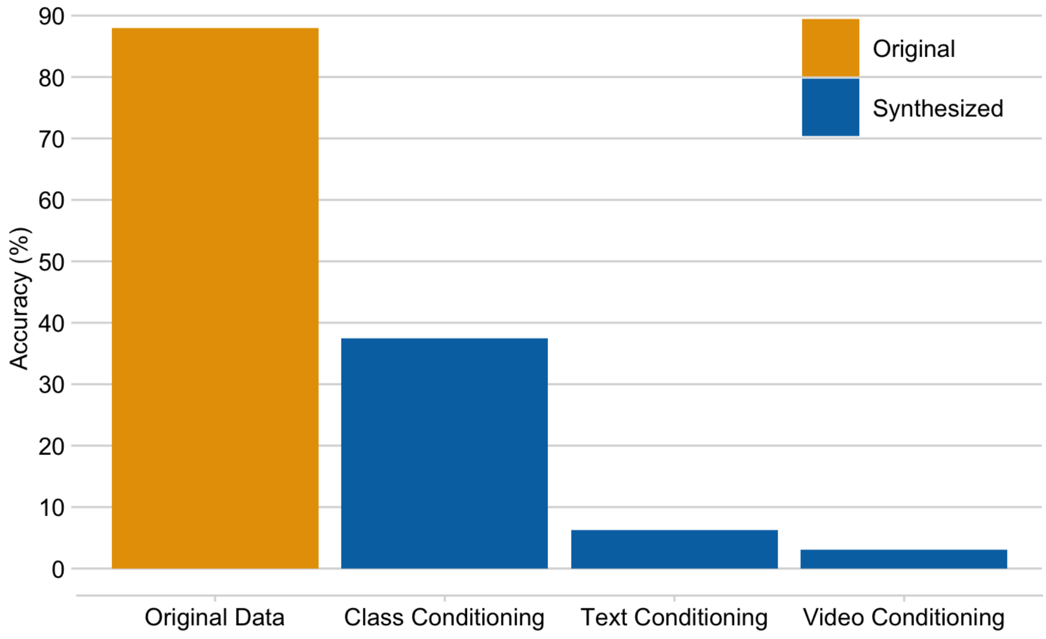

Figure 3: Speech to Text Accuracy

on 3136 by 768 video embeddings. The class conditioning and text conditioning models constitute intermediate models that serve as milestones towards the video model. The correct label can be ascertained by human listeners upon inspecting the generated audio of the class-conditioned model, suggesting that class-conditioning the sampling process on a 512-dimensional projection of a 39 dimensional vector works relatively well. As can be observed in Figure 3, the outputs of the Class Conditioning Model achieve 38% accuracy when fed to the Speech-to-Text Whisper model. The training loss of the Class Conditioned model is presented in Figure 4.

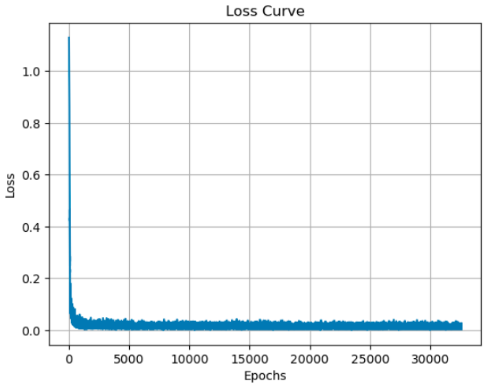

Figure 4: Class Conditioning Train Loss

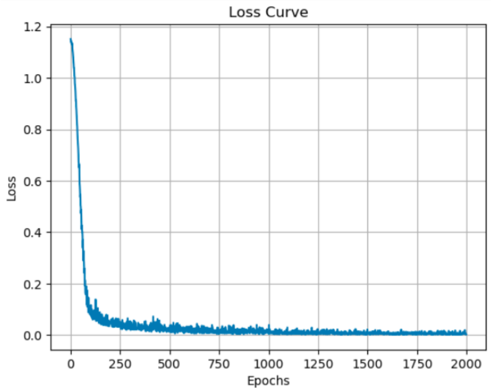

Figure 5: Video Conditioning Train Loss

The text-conditioned model performs audibly worse than the class-conditioned model, which is corroborated by the 6.3% Speech-to-Text accuracy shown in Figure 3. This considerable decrease in performance can be attributed to two main reasons: first, because a single linear projection layer that takes in 768 dimensional tensors and outputs a 512 dimensional tensor might not contain enough nonlinearities to disentangle word vectors that are close together in word-embedding space; and secondly, due to the fact that we are projecting vectors down from 768 to 512 as opposed to projecting upwards from 39 to 768, potentially leading to catastrophic data loss.

The video-conditioned model exacerbates the two aforementioned problems: the linear transformation carried out by the projection layer takes in 2,408,448 floating-point values and reduces it to 512 values in the video setting, undoubtedly leading to signal loss from the video information, and likely not having sufficient expressiveness to efficiently encode the video features in 512 dimensions. Although the audio samples generated by the model conditioned on video embeddings sound like natural speech, the generated audio does not relate to the audio from the videos they were conditioned on: the Whisper Speech to Text model correctly identified 3.1% of generated samples' text label.

The Speech-to-Text transcription accuracy model rankings are corroborated by another evaluation metric: the distribution of the distance between the generated samples and the true class samples.

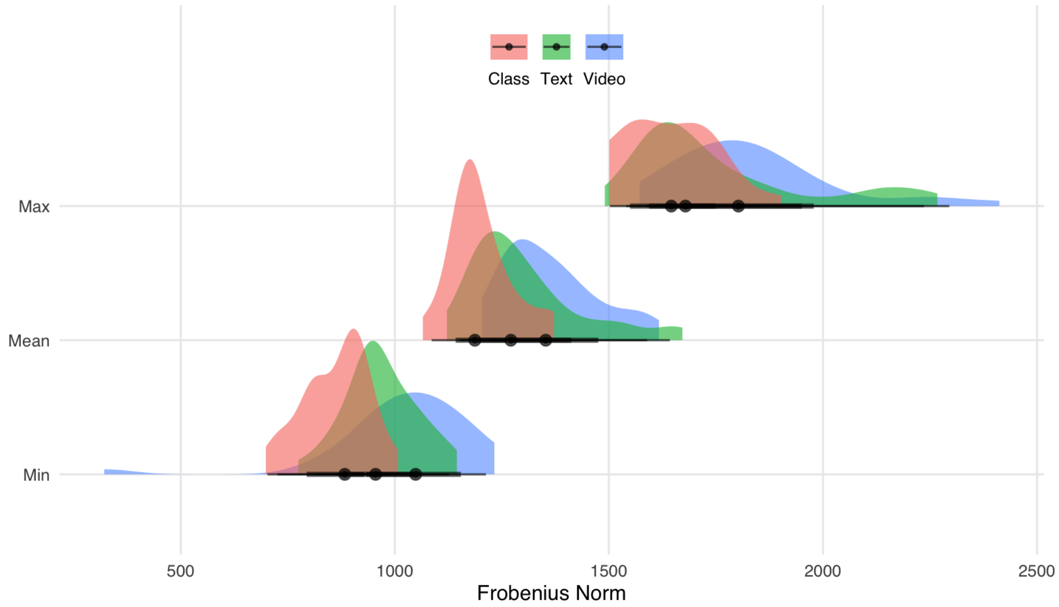

Figure 6: Distribution of Frobenius Norm

As can be observed in Figure 6 above, the class-conditioned model shown in red yields samples that are in expectation closer to the true samples in their corresponding class than the text-conditioned and video-conditioned model, with the video-conditioned model yielding worst results. This ranking stays consistent across all three statistics: the minimum, which is the distribution of distance to the nearest neighbor; the mean distance; and the maximum distance.

4 Limitations

This project does encounter several limitations to consider.

We utilized a standard diffusion model in the Melspectrogram space, but an alternative could have been employing a latent diffusion model. Latent diffusion models offer potential benefits such as improved efficiency and potentially better handling of the high-dimensional data typical of video and audio inputs. The exploration of a latent diffusion approach may offer insights into more effective or efficient synthesis processes.

In terms of video feature extraction, our project solely utilized the encoder from a Masked Autoencoder (MAE), and projected the embedding using a single linear embedding layer. The method fell short in capturing the nuanced audio-visual correlations critical for high-quality speech synthesis. Exploring other models like the Contrastive Audio-Visual Pretrained (CAVP) model could potentially enhance our approach by providing a more robust and detailed representation of audio-visual features [see for examplee Chen et al., 2020]. Additionally, appending more layers between the embedding and U-Net's timestep embedding could aid in more efficiently encoding the relevant audio-visual features into the timestep embedding space.

One significant limitation of our study is the absence of an objective metric for audio, or more specifically speech, quality. In this direction, we propose two proxy measurements that are aligned with the quality of our samples; however they do not measure directly sound quality.

5 Future Directions

As mentioned in section 4, the immediate next steps would be to integrate a latent diffusion model, explore more advanced video encoding techniques and include benchmarking against established methods.

In terms of application, finetuning the model on individual speakers could significantly enhance the personalization and realism of synthesized speech. The system could learn unique vocal characteristics, such as pitch, tone, and speaking style, ensuring that the synthesized audio not only matches the lip movements but also reflects the speaker's distinct vocal traits. Moreover, advancing towards real-time processing would enable applications in live environments, such as real-time audio synthesis for mute video calls or live broadcast dubbing.

Another possibility would be to incorporate emotional intelligence into the models to detect and replicate the emotional state of speakers. This could improve the naturalness of synthesized speech and involves analyzing facial expressions and body language in addition to lip movements to infer and generate speech with corresponding emotional tones.

As the technology progresses, it is crucial to address ethical concerns related to privacy and consent, as well as mitigate any biases related to language, dialect, or accent to ensure fairness and inclusiveness.

In broader terms, our conditional diffusion approach to generating other modalities can be expanded further and transferred to many different contexts. To do so, multimodal conditioning ca

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. 2023. doi: https://doi.org/10.48550/arXiv.2306.17203.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *SciPy*, 2015. URL `https://api.semanticscholar.org/CorpusID:33504`.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C V Jawahar. Learning individual speaking styles for accurate lip to speech synthesis, 2020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022.