

DiffV2S: Diffusion-based Video-to-Speech Synthesis with Vision-guided Speaker Embedding

Jeongsoo Choi* Joanna Hong* Yong Man Ro
School of Electrical Engineering, KAIST
{jeongsoo.choi, joanna2587, ymro}@kaist.ac.kr

Abstract

Recent research has demonstrated impressive results in video-to-speech synthesis which involves reconstructing speech solely from visual input. However, previous works have struggled to accurately synthesize speech due to a lack of sufficient guidance for the model to infer the correct content with the appropriate sound. To resolve the issue, they have adopted an extra speaker embedding as a speaking style guidance from a reference auditory information. Nevertheless, it is not always possible to obtain the audio information from the corresponding video input, especially during the inference time. In this paper, we present a novel vision-guided speaker embedding extractor using a self-supervised pre-trained model and prompt tuning technique. In doing so, the rich speaker embedding information can be produced solely from input visual information, and the extra audio information is not necessary during the inference time. Using the extracted vision-guided speaker embedding representations, we further develop a diffusion-based video-to-speech synthesis model, so called DiffV2S, conditioned on those speaker embeddings and the visual representation extracted from the input video. The proposed DiffV2S not only maintains phoneme details contained in the input video frames, but also creates a highly intelligible mel-spectrogram in which the speaker identities of the multiple speakers are all preserved. Our experimental results show that DiffV2S achieves the state-of-the-art performance compared to the previous video-to-speech synthesis technique.

1. Introduction

Video-to-speech synthesis techniques [7, 15, 21, 22, 23, 33, 34, 39] have been broadly studied in lip-reading research areas. It reconstructs speech from a silent talking face video, which has an advantage of not requiring extra text information of the given video input during training. However, it is still regarded as a challenging task, especially in multi-

speaker and noisy environment settings, since the video-to-speech synthesis technique needs to capture the complex relationship between various lip movements and speech.

The relationship between lip movements and speech is not always straightforward; there is considerable variation in how different people articulate sounds, as well as how their lip movements are affected by factors such as facial expressions, accents, and noise. To resolve the complicated factors that speakers themselves contain, several recent studies [7, 23, 33, 39] have utilized extra speaker embedding representations obtained from the original audio information of the video input with the same speaker. The speaker embeddings are helpful for obtaining the speaker's characteristics, where those characteristics cannot be directly derived from silent talking face video. However, directly manipulating the reference audio information during the inference is not always possible as the audio information is sometimes unobtainable because of noisy environments, absence of speech, and unseen speakers during the inference time.

To alleviate the aforementioned issue, we present a novel vision-guided speaker embedding extractor using a self-supervised pre-trained model. With the largely trained audio-visual speech representation model [40], we adopt a prompt tuning technique [28] to train the certain part of the model in order to extract the appropriate speaker embedding features from the input video sequences. We set only a small amount of downstream task-specific parameters as the learnable parameters for extracting speaker embedding into the input space while freezing the other parts of the pre-trained model. By doing so, the rich speaker embedding information can be produced from solely on input visual information, and the extra audio information is not necessary during the inference time period.

Furthermore, we propose a conditional diffusion-based video-to-speech synthesis model, named DiffV2S, using the vision-guided speaker embedding representations. As the denoising diffusion model has been proven to be effective in generating semantically meaningful representation in both image and audio processing [18, 32, 43], we also newly adopt the diffusion model to achieve high-quality mel-

*Both authors have contributed equally to this work.

spectrogram containing semantically detailed information. The proposed DiffV2S is comprised of conditional diffusion modeling and sampling with speaker embedding guidance. During training, the proposed DiffV2S reconstructs a mel-spectrogram from a standard Gaussian distribution with the condition of speaker embedding representations concatenated with the visual features extracted from the input silent talking face video. During sampling, the speaker characteristics are driven to enable the model to properly adopt the speaker’s style, such as voice and accent, while maintaining the articulate phoneme details faithfully. Therefore, our model not only maintains phoneme details contained in the input video frames, but also creates a noise-free and highly intelligible mel-spectrogram in which the speaker identity characteristics are entirely preserved.

To validate the effectiveness of the proposed method, we utilize LRS2 [9] and LRS3 [1], the largest sentence-level audio-visual datasets obtained in the wild. Through comprehensive experiments, we show that the generated speech from the proposed DiffV2S contains much detailed contents, thus producing noise-free audio waveform with high performances.

Our key contributions are as follows:

- We propose a vision-guided speaker embedding extractor, so the rich speaker information can be produced solely from the input video frames. In doing so, the audio information is not necessary during the inference time.
- We present the novel diffusion-based video-to-speech synthesis model, DiffV2S, conditioned on the speaker embedding representations and the visual representation extracted from the input talking face video. The DiffV2S not only maintains phoneme details contained in the input video frames, but also creates a highly intelligible mel-spectrogram in which the speaker identities are all preserved.
- To best of our knowledge, this is first time to utilize the diffusion model in video-to-speech synthesis. The generated speech from the proposed DiffV2S contains much detailed information, thus producing noise-free audio waveform with high performances.

2. Related Works

2.1. Video-to-Speech Synthesis

Video-to-speech synthesis is one of the lip-reading techniques that have been consistently studied. Ephrat and Peleg [12] presented the initial deep-learning based video-to-speech method using an end-to-end CNN-based model. Akbari *et al.* [2] utilized autoencoders in presenting a reconstruction-based video-to-speech synthesis. Prajwal

et al. [39] introduced Lip2Wav model using a well-known sequence-to-sequence architecture to correctly capture the context. Hong *et al.* [15] adopted a multi-modal memory network in video-to-speech synthesis to associate an extra audio information during inference time period. GAN-based techniques [16, 22, 34, 45] were presented to produce realistic utterances from the silent talking face videos. Recent video-to-speech techniques [7, 23, 33] utilized extra speaker embeddings from the original audio information in order to obtain the speaker’s speaking styles and characteristics. Instead of directly using the audio information, in this work, we try to extract the speaker information from the input silent talking face video sequences. To do so, we adopt prompt tuning technique.

2.2. Prompt Tuning

Prompt tuning refers to the process of adjusting a pre-trained model by giving it additional prompts that are relevant to a specific task or domain. With a great development of the large pre-trained language model like GPT-3 [3], prompt tuning has been drawn attention in utilizing a frozen language model for a specific task, reducing the number of model parameters for training and memory usage. Liu *et al.* [31] firstly introduced P-tuning to add trainable continuous embeddings, so called continuous prompts, to the original sequence of input word embeddings. Lester *et al.* [28] presented the modification of initial P-tuning technique by simply prepending the prompt to the input. Zhong *et al.* [47] designed an effective continuous method for optimizing prompt, so called OptiPrompt. Liu *et al.* [30] further generalized P-tuning so that it can be comparable to fine-tuning universally across various model scales and natural language understanding tasks. While P-tuning only focuses on language models, Jia *et al.* [19] proposed a visual prompt tuning technique to fine-tune for large-scale Transformer models in vision. Inspired by [19, 28], in this paper, we adopt prompt tuning technique to the large pre-trained audio-visual representation model in order to obtain a proper speaker representation from the visual input.

2.3. Diffusion Model

Diffusion model has been spotlighted in many research areas regarding image generation. Sohl-Dickstein *et al.* [41] firstly introduced diffusion probabilistic models which is a parameterized Markov chain trained using variational inference to generate samples that match the data within a specified time. Ho *et al.* [14] presented progress of diffusion models that they are capable of generating high quality image samples. Along with the large usage of diffusion models in image-video generation areas, there also have been numerous studies in synthesizing audio using diffusion models. Diffusion based neural vocoders [5, 6, 26] were proposed to model the fine details of waveform conditioned on

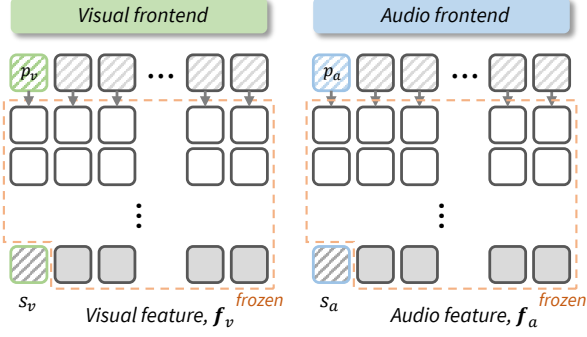


Figure 1. Prompt tuning via self-supervised audio-visual pre-trained model.

mel-spectrogram. Several text-to-speech (TTS) techniques [18, 29, 38] also utilized the diffusion models and achieved high-fidelity and efficient speech synthesis. While the denoising diffusion model has been proven to be effective in generating semantically meaningful representation in both image and audio processing, there has not yet been explored in the video-to-speech study. Thus, it is the first time to utilize the diffusion probabilistic model in video-to-speech technique.

3. Methodology

Given the input talking face video sequence, $\mathbf{x} = \{x_1, \dots, x_L\} \in \mathbb{R}^{L \times H \times W \times C}$ where L , H , W , and C are the frame length, height, width, and channel sizes, respectively, we design a model that synthesizes the proper mel-spectrogram, $\mathbf{M} = \{M_1, \dots, M_S\} \in \mathbb{R}^{K \times S}$ with K mel-spectral channel and the sequence length S . The main goal of this paper is to reconstruct the mel-spectrogram which contains the right styles of the input speaker as well as the articulate phoneme details. To this end, we firstly propose a vision-guided speaker embedding extractor composed of a largely pre-trained audio-visual speech representation model modified by prompt tuning. Using the speaker embedding features, we also design a diffusion based speech synthesis model conditioned on the extracted speaker embedding. We will explain the detailed aforementioned techniques in the following subsections.

3.1. Vision-guided Speaker Embedding Extractor

When synthesizing the speech from a silent talking face video, it is important to know speaker’s characteristics, tones, and accents to represent the proper acoustic sound. It is not always possible to extract the speaker embedding features from the audio information during the inference time due to the absence of speech and noisy environment. Thus, we design a vision-guided speaker embedding extractor that can generate the adequate speaker embedding feature without any additional audio guidance. To do so, we adopt prompt tuning [28] technique which leverages few continuous learn-

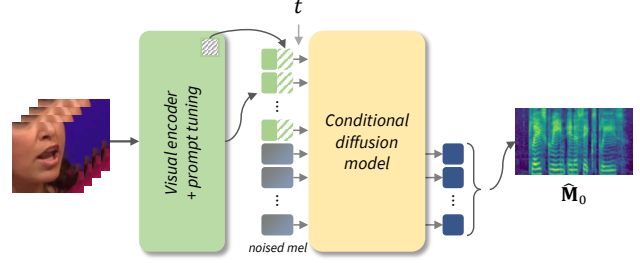


Figure 2. Training procedure of the proposed speaker embedding conditioned diffusion model.

able parameters in order to serve as prompts fed into a largely pre-trained audio-visual representation model. The prompt tuning technique is beneficial for aggregating every informative representation from the pre-trained model with adjusting the small amount of parameters for the downstream task, thus regarded as cost-effective and highly-informative.

3.1.1 Prompt Tuning for Speaker Embedding Extractor

Given a large pre-trained audio-visual representation model, we set d -dimensional learnable parameters $p_v \in \mathbb{R}^d$ and $p_a \in \mathbb{R}^d$, also called prompts, for the input video \mathbf{x} and the mel \mathbf{M} , respectively. The prompts p_v and p_a are trained for extracting the speaker embeddings. To do so, we firstly extract the visual and audio embeddings, $\mathbf{e}_v \in \mathbb{R}^{L \times d}$ and $\mathbf{e}_a \in \mathbb{R}^{L \times d}$, from the visual and audio frontends, respectively:

$$\mathbf{e}_v = \mathcal{F}_v(\mathbf{x}) \quad (1)$$

$$\mathbf{e}_a = \mathcal{F}_a(\mathbf{M}). \quad (2)$$

Then, each of the extracted visual and audio embeddings are taken into each pre-trained visual and audio feature extractor, Φ_v and Φ_a respectively, where the prompt is added:

$$[k_{v,1}, \mathbf{e}_{v,1}] = \Phi_{v,1}(p_v, \mathbf{e}_{v,0}) \quad (3)$$

$$[k_{v,i+1}, \mathbf{e}_{v,i+1}] = \Phi_{v,i+1}(k_{v,i}, \mathbf{e}_{v,i}), \quad (4)$$

for the i -th layer where $i = 1, 2, \dots, n - 1$. We design the vision prompt p_v to affect only the final layer $k_{v,n}$ without updating any other features through the self-attention mask. Thus, the last layer embedding $k_{v,n}$ becomes the vision-guided speaker embedding $s_v \in \mathbb{R}^d$, and the remainders $\mathbf{e}_{v,n}$ becomes $\mathbf{f}_v \in \mathbb{R}^{L \times d}$. Same procedures are applied for the audio embeddings \mathbf{e}_a . During the entire training phase, we only train the learnable prompts, p_v and p_a . We keep the entire backbone pre-trained model, audio and visual encoder which of each contains the combination of the frontend and the feature extractor for the corresponding modality, remaining frozen. We also utilize the self-attention mask so that the additional prompt learns from the layer-wise feature of the pre-trained model but does not affect to the original features.

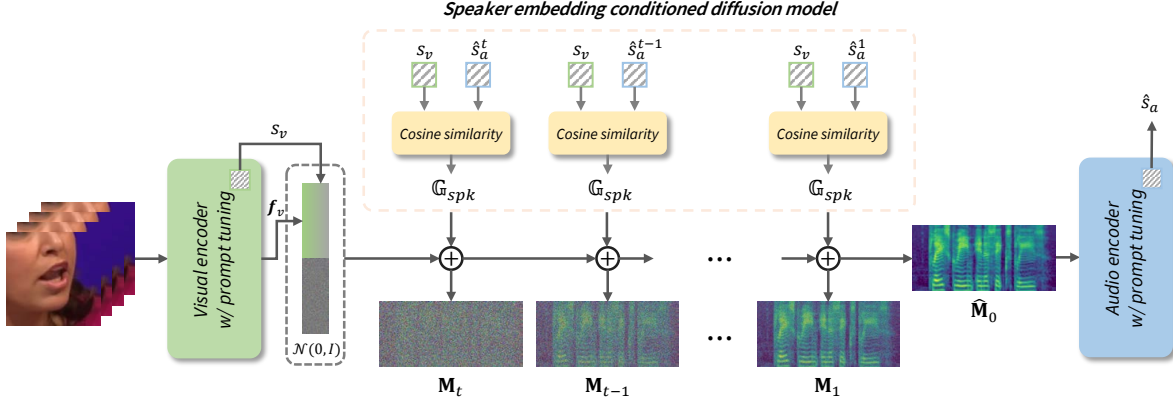


Figure 3. Sampling procedure of the proposed speaker embedding conditioned diffusion model.

The visualization of the prompt tuning technique is indicated in Figure 1.

3.1.2 Training Speaker Embedding Extractor

In order to train the learnable parameters, p_v and p_a , to correctly extract the speaker embeddings, we adopt a pre-trained speaker encoder [46] as a ground-truth guidance of speaker embeddings extracted from the actual audio input during training. It is originally trained for speaker verification task on a combination of the large speech datasets [8, 35, 37], which is trained to optimize a generalized speaker verification loss. We now call s_G as the speaker embedding guidance extracted from [46].

To begin with, we set each s_v and s_a to be similar to the speaker embedding guidance s_G . Given a batch of N pairs of vision-guided speaker embedding s_v and the speaker embedding guidance s_G , the prompt p_v is trained to predict which of the $N \times N$ possible pairings across a batch actually occurred. We utilize InfoNCE loss [36] to encourage the prompts p_v and p_a to place the representations of positive pairs close to each other and the representations of negative pairs far apart:

$$\mathcal{L}_{s_v, s_G} = -\log \frac{\exp(s_{v,i} \cdot s_{G,j}) / \tau}{\sum_{k=0}^N \exp(s_{v,i} \cdot s_{G,k}) / \tau}. \quad (5)$$

Here, we fix the updates of the speaker embedding guidance s_G . Likewise, \mathcal{L}_{s_a, s_G} can be obtained with above equation. Furthermore, we guide both s_v and s_a to be mapped into common embedding space through \mathcal{L}_{s_a, s_v} and \mathcal{L}_{s_v, s_a} so that both embeddings can relate. The final loss function of training the speaker embedding extractor is following:

$$\mathcal{L}_{spk_emb} = \mathcal{L}_{s_v, s_G} + \mathcal{L}_{s_a, s_G} + \mathcal{L}_{s_v, s_a} + \mathcal{L}_{s_a, s_v}. \quad (6)$$

Therefore, given a new video sequence or mel sequence, the speaker embedding extractor can retrieve the most plausible speaker representations from the joint embedding, even if they are not seen during training.

3.2. Diffusion-based Video-to-Speech Model

3.2.1 Training Procedure of DiffV2S

The proposed DiffV2S generates a detailed mel-spectrogram from a standard Gaussian distribution conditioned on the channel-wise concatenation of the vision-guided speaker embeddings s_v and the visual features f_v :

$$c = f_v || s_v. \quad (7)$$

Let $\mathbf{M}_1, \dots, \mathbf{M}_T$ be a sequence of variables with T number of timesteps. The forward diffusion process transforms mel-spectrogram \mathbf{M}_0 into a Gaussian noise \mathbf{M}_T through Markov chain transitions with a predefined variance schedule β_t :

$$q(\mathbf{M}_t | \mathbf{M}_{t-1}, c) = \mathcal{N}(\mathbf{M}_t; \sqrt{1 - \beta_t} \mathbf{M}_{t-1}, \beta_t \mathbf{I}). \quad (8)$$

The reverse process is a backward of forward diffusion process, where it recovers a mel-spectrogram from a standard Gaussian noise. It can be defined as the conditional distribution and factorized into multiple transitions based on Markov chain property:

$$p_\theta(\mathbf{M}_{0:T}, c) = p(\mathbf{M}_T, c) \prod_{t=1}^T p_\theta(\mathbf{M}_{t-1} | \mathbf{M}_t, c), \quad (9)$$

where

$$p_\theta(\mathbf{M}_{t-1} | \mathbf{M}_t, c) = \mathcal{N}(\mathbf{M}_{t-1}; \mu_\theta(\mathbf{M}_t, t, c), \sigma_\theta(\mathbf{M}_t, t, c) \mathbf{I}). \quad (10)$$

Here, μ_θ and σ_θ are the mean and variance for the denoising model. In order to predict $q(\mathbf{M}_0 | c)$, we need to optimize the negative log-likelihood of the predicted mel-spectrogram: $\mathbb{E}_q[\log p_\theta(\mathbf{M}_0 | c)]$. Since $p_\theta(x_0 | c)$ is intractable, the reparameterization trick [14] is demonstrated to calculate the variational lower bound of the log-likelihood in a closed form. In this case, the model learns to find $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Nonetheless, instead of modelling $\epsilon_\theta(\mathbf{M}_t, t)$, the proposed DiffV2S is designed to predict the mel-spectrogram itself,

Algorithm 1 Diffusion Sampling Procedure of DiffV2S with Vision-guided Speaker Embedding

- 1: **Inputs:** source talking face video sequence \mathbf{x} , target mel-spectrogram $\mathbf{M}_{\text{target}}$, learnable prompts p_v and p_a
 - 2: **Outputs:** synthesized mel-spectrogram $\hat{\mathbf{M}}_0$
 - 3: $\mathbf{M}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 - 4: $\mathbf{e}_v \leftarrow \mathcal{F}_v(\mathbf{x})$
 - 5: $[s_v, \mathbf{f}_v] \leftarrow \Phi_v(p_v, \mathbf{e}_v)$
 - 6: $c \leftarrow \mathbf{f}_v \| s_v$
 - 7: **for all** t from T to **1 do**
 - 8: $\hat{\mathbf{M}}_0 \leftarrow \Psi_\theta(\mathbf{M}_t, t, c)$
 - 9: $\hat{\mathbf{e}}_a \leftarrow \mathcal{F}_a(\hat{\mathbf{M}}_0)$
 - 10: $[\hat{s}_a, \hat{\mathbf{f}}_a] \leftarrow \Phi_a(p_a, \hat{\mathbf{e}}_a)$
 - 11: $\mathbb{G}_{spk} \leftarrow (1 - \text{sim}(s_v, \hat{s}_a))$
 - 12: $\hat{\epsilon} \leftarrow \frac{\mathbf{M}_t - \sqrt{\bar{\alpha}_t} \hat{\mathbf{M}}_0}{\sqrt{1 - \bar{\alpha}_t}} - \sqrt{1 - \bar{\alpha}_t} \nabla_{\mathbf{M}_t} \lambda \mathbb{G}_{spk}$
 - 13: $\mathbf{M}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{M}_t - \sqrt{1 - \bar{\alpha}_t} \hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \hat{\epsilon}$
 - 14: **end for**
 - 15: **return** $\hat{\mathbf{M}}_0$
-

as shown in Figure 2. Thus, the diffusion model $\Psi(\mathbf{M}_t, t, c)$ predicts the mel-spectrogram $\hat{\mathbf{M}}_0$, and we utilize the L1 reconstruction loss between the predicted one and the ground truth one as follows:

$$\mathcal{L}_{diff} = \mathbb{E}_{\mathbf{M}_0 \sim q(\mathbf{M}_0|c)} [\|\mathbf{M}_0 - \Psi_\theta(\mathbf{M}_t, t, c)\|_1]. \quad (11)$$

3.2.2 Conditional Sampling for DiffV2S

During the sampling procedure, we utilize the vision-guided speaker embeddings s_v to incorporate the gradients from the guidance of speaker embedding representations as a condition in order to guide the model to sample the mel-spectrogram with the desired speaker characteristics. To do so, inspired by [20], we firstly formulate the cosine similarity loss between the vision-guided speaker embedding s_v and the audio-guided speaker embedding \hat{s}_a extracted from the predicted mel-spectrogram $\hat{\mathbf{M}}_0$:

$$[s_v, \mathbf{f}_v] = \Phi_v(p_v, \mathcal{F}_v(\mathbf{x})), \quad (12)$$

$$[\hat{s}_a, \hat{\mathbf{f}}_a] = \Phi_v(p_a, \mathcal{F}_a(\hat{\mathbf{M}}_0)), \quad (13)$$

$$\mathbb{G}_{spk} = 1 - \text{sim}(s_v, \hat{s}_a), \quad (14)$$

where sim corresponds to cosine similarity.

Using the gradients of \mathbb{G}_{spk} with respect to \mathbf{M}_t , it is possible to derive the conditional sampling with the score function following [11]. We utilize the deterministic sampling method DDIM [42] for enhancing the sampling speed. We derive the epsilon prediction first; then, we calculate \mathbf{M}_{t-1} as follows:

$$\hat{\epsilon} = \frac{\mathbf{M}_t - \sqrt{\bar{\alpha}_t} \hat{\mathbf{M}}_0}{\sqrt{1 - \bar{\alpha}_t}} - \sqrt{1 - \bar{\alpha}_t} \nabla_{\mathbf{M}_t} \lambda \mathbb{G}_{spk}, \quad (15)$$

$$\mathbf{M}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{M}_t - \sqrt{1 - \bar{\alpha}_t} \hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \hat{\epsilon}, \quad (16)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. Algorithm 1 presents the more details of sampling algorithm.

4. Experimental Setup

4.1. Datasets

LRS2-BBC [9] is an English sentence-level audio-visual dataset, collected from BBC television shows. It contains more than 2000 hours of videos, where both pre-train and train sets consist of about 142,000 utterances, validation set includes about 1,100 utterances, and test set contains 1,200 utterances. We utilize both the pre-training and training sets for the training, and the test set for the inference.

LRS3-TED [1] is also an audio-visual dataset in English, collected from TED and TEDx videos. It includes unconstrained long sentences with more than 50,000 vocabularies and thousands of speakers. It contains about 150,000 videos which are total about 439 hours long. About 131,000 utterances are utilized for training, and about 1,300 utterances are used for testing. We follow the unseen data splits of [33].

4.2. Implementation Details

4.2.1 Data Preprocessing

For every video, we crop based on the lip-centered region and resize the image into 88×88 . We convert the mel-spectrogram from 16kHz audio using the filter size of 640 and hop size of 160 with 80 mel bands which becomes a sampling rate of 100Hz. In order to match the length of the audio feature with the visual feature of 25Hz, 40ms of mel-spectrogram is stacked to make an audio feature with dimension of 320. The mel-spectrogram is converted into a log scale and normalized into $[-1, 1]$ before entering our model.

4.2.2 Architectural Details

For the visual encoder, the visual features with dimension of 1024 are obtained by using publicly available LARGE AV-HuBERT [40], and the speaker embedding is obtained through prompt tuning with the form of a vector with 1024 dimension followed by the linear layer becoming a size of 256. The speaker embedding is repeated in every frame to match the length of the visual feature sequence. We use a linear layer to project the concatenated visual features and speaker embedding to dimension of 512. The same procedure applies for the audio encoder and prompt tuning for audio representation. During training the vision-guided speaker embedding extractor through prompt tuning technique, we utilize the self-attention mask so that the additional prompt learns from the layer-wise feature of the pre-trained model but does not

Method	LRS2-BBC					LRS3-TED				
	Low-level		Synchronization		Content	Low-level		Synchronization		Content
	ESTOI \uparrow	MCD \downarrow	LSE-C \uparrow	LSE-D \downarrow	WER \downarrow	ESTOI \uparrow	MCD \downarrow	LSE-C \uparrow	LSE-D \downarrow	WER \downarrow
<i>with speaker embedding from audio</i>										
SVTS [33]	0.331	6.86	7.80	6.47	71.4%	0.271	8.02	6.04	8.28	78.0%
Multi-task [23]	0.341	9.37	6.88	7.32	57.8%	0.268	9.89	5.19	8.89	65.8%
<i>without speaker embedding from audio</i>										
VCA-GAN [22]	0.134	9.35	2.63	11.61	101.1%	0.207	8.85	4.54	9.63	95.9%
SVTS [33]	0.301	7.97	7.87	6.30	76.6%	0.244	8.60	7.08	7.04	81.9%
Multi-task [23]	0.322	10.22	7.19	7.01	61.0%	0.240	10.16	4.85	9.15	74.8%
Proposed model	0.283	9.85	7.51	6.90	52.7%	0.284	9.35	7.28	7.27	39.2%

Table 1. Performance comparisons on LRS2 and LRS3 datasets. \uparrow means that the higher is the better, and \downarrow means that the lower is the better.

affect to the original features. For the diffusion model, we adopt Transformer [44] encoder architecture as a training network. We use 8 Transformer layers with 4 attention heads, hidden dimension of 512, and feed-forward layer dimension of 1024. The model uses GELU [13] activation. The sinusoidal positional embeddings are added to each audio and visual features which means temporally synced audiovisual feature share the same positional embedding. All the input features are linearly projected into 512 dimensions before taken into the diffusion model. Lastly, we utilize HiFi-GAN [25] neural vocoder to convert sampled mel-spectrogram into the actual waveform output.

4.2.3 Training Details

For training, we use AdamW [24] optimizer with learning rate of 10^{-4} . We set 1000 timesteps ($T = 1000$) as default for sampling the diffusion backward process and $\lambda = 1000$ to determine the guidance level of the input during sampling. We train DiffV2S for 300k updates on a GPU with batch size of 64. For computing, we use a single A6000 GPU. Note that for VCA-GAN [22], SVTS [33] (LRS3 dataset only), and Multi-task [23], we are provided with the test audio samples from the authors. Otherwise, we re-implement and train the previous works, generate the test audio samples, and evaluate the performances.

4.3. Evaluation Metrics

To measure the low-level quality of the generated waveform, we utilize Extended STOI (ESTOI) [17] as a measurement of the intelligibility of the generated speech and Mel-cepstral distortion (MCD) [27] which quantifies the distance between the generated audio signal and the ground truth audio signal with mel-frequency cepstrum, focusing more on details. In Addition, we verify the synchronization compared to the ground truth speech. We adopt SyncNet [10] and predict the temporal distance between audio and video (LSE-D) and the prediction’s average confidence (LSE-C). We utilize Word Error Rate (WER) in order to evaluate the content quality of the generated speech. We also measure

Method	Naturalness	Intelligibility	Voice matching
<i>with spk-emb from audio</i>			
SVTS [33]	2.16 \pm 0.24	2.50 \pm 0.28	2.15 \pm 0.25
Multi-task [23]	1.83 \pm 0.28	2.37 \pm 0.33	1.96 \pm 0.31
<i>without spk-emb from audio</i>			
VCA-GAN [22]	1.35 \pm 0.12	1.76 \pm 0.31	1.39 \pm 0.15
SVTS [33]	1.80 \pm 0.25	2.35 \pm 0.27	1.74 \pm 0.21
Multi-task [23]	1.53 \pm 0.18	2.19 \pm 0.34	1.60 \pm 0.21
Proposed model	4.68\pm0.18	3.59\pm0.18	3.91\pm0.27
Actual Voice	4.98 \pm 0.02	4.93 \pm 0.03	-

Table 2. MOS comparison of the audio samples of LRS3 dataset.

the Speaker Encoder Cosine Similarity (SECS) [4] between speaker embeddings of the audio samples extracted from the speaker encoder [46] in order to verify how much the generated speech and the original speech have similar speaker voice. Lastly, we conduct a human subjective study through mean opinion scores (MOS) of naturalness, intelligibility, and voice matching of the generated speech. Note that we focus on WER, SECS, and MOS metrics to examine whether the proposed model generates the clean speech samples with the right content and voice.

5. Experimental Results

5.1. Quantitative Results

5.1.1 Comparisons with the State-of-the-Arts

To begin with, we compare the performances with the previous methods [22, 23, 33] using LRS2-BBC and LRS3-TED datasets. Since SVTS [33] and Multi-task [23] utilize the audio guidance for extracting the speaker embedding, we reproduce both framework disregarding the audio guidance during the inference time and name *without speaker embedding from audio* following the name of the work, as shown in Table 1. Note that we also name the future tables in the following way.

The proposed model well synchronizes with the actual input video, obtaining 7.51 LSE-C and 6.90 LSE-D on LRS2

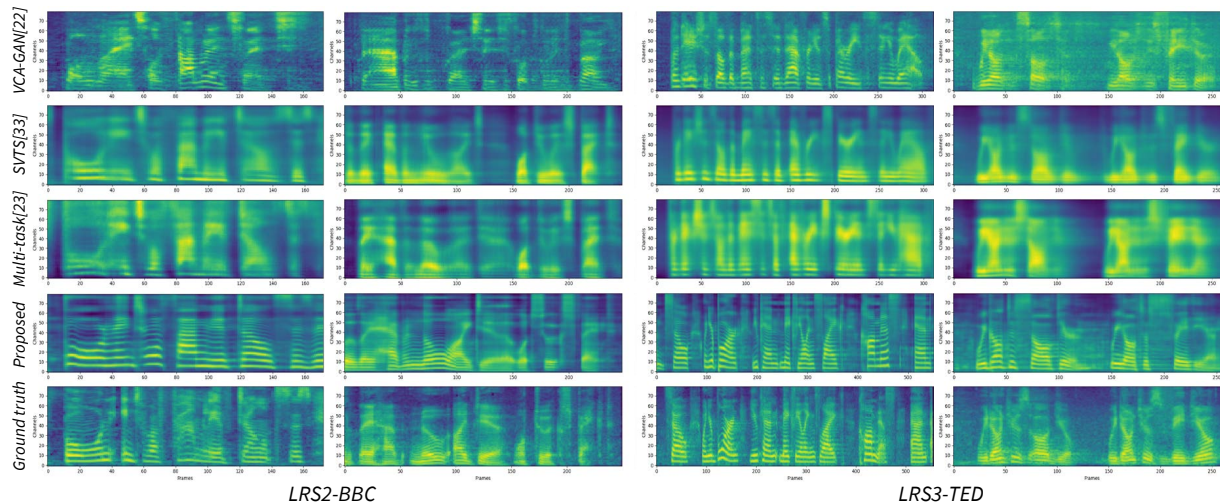


Figure 4. The sample mel-spectrogram visualizations on LRS2-BBC and LRS3-TED datasets from the previous methods [22, 23, 33], the proposed model, and the ground truth.

and 7.28 LSE-C and 7.27 LSE-D on LRS3 dataset. On the low-level quality criteria, the model attains 0.283 ESTOI and 9.85 MCD on LRS2 and achieves 0.284 ESTOI and 9.35 MCD for LRS3 dataset, which are comparable scores with other methods. Most importantly, the DiffV2S considerably outperforms in regarding content quality part, achieving 52.3% and 39.2% on LRS2 and LRS3, respectively, which are 5.1% and 26.6% WER performance gaps, on both LRS2 and LRS3 respectively, compared to Multi-task [23] which best performs on content quality. Note that we focus more on the content quality extracted from the output generated waveform. Therefore, we can infer that the detailed generated mel-spectrogram also lead to the audio speech reconstruction which contain clear and right content information.

5.1.2 Human Subjective Study

In order to evaluate the generated speech quality, we additionally conduct a human subjective study through mean opinion scores (MOS) via three objective metrics: *naturalness*, *intelligibility*, and *voice matching*. For the *naturalness*, it verifies whether the generated speech is natural, similar to actual human speech, and the *intelligibility* evaluates whether the words in the generated speech clearly sound compared to the actual transcription. Lastly, the *voice matching* determines how well the results of the proposed model matches the voice of the target speaker. We use 20 randomly sampled audio from the test dataset of LRS3-TED, where each sample is evaluated by 15 participants in a 5-point scale with 0.5 increment. We also measure the 95% confidence intervals of the total score of the participants.

Table 2 indicates the MOS comparisons of the audio samples from the previous methods [22, 23, 33], the proposed method, and the ground truth audio samples from LRS3-TED dataset. The proposed model achieves 4.68 MOS for

Method	LRS2	LRS3
<i>with spk-emb from audio</i>		
SVTS [33]	0.558	0.623
Multi-task [23]	0.525	0.549
<i>without spk-emb from audio</i>		
VCA-GAN [22]	0.453	0.445
SVTS [33]	0.499	0.543
Multi-task [23]	0.457	0.495
Proposed model	0.581	0.625

Table 3. Speaker Encoder Cosine Similarity (SECS) between the speaker embeddings of the generated audio samples and the ground truth audio samples.

naturalness and 3.59 MOS for *intelligibility*. It is shown that there are large performance gaps between the MOS of the proposed framework and those of the other methods, meaning that the audio samples from the proposed model contain noise-free and detailed audio waveform. Next, for *voice matching* criterion, it is clearly shown that the audio samples generated from the proposed model best follow the actual voice characteristics, achieving 3.91 MOS. This verifies that the vision-guided speaker embedding representations are well extracted from the input videos and are adequately conditioned in the mel-spectrogram generation process, thus producing the voices similar to the ground truth voices. In addition, if the audio-guide is absent, the scores of not only *voice matching* criterion but also *naturalness* and *intelligibility* get dropped. The means that the audio-guide actually helps the network producing proper mel-spectrogram to be similar to the actual ground truth one.

5.1.3 Vision-guided Speaker Embedding Analysis

Lastly, we analyze the vision-guided speaker embedding to verify that the extracted speaker embedding from the

Baseline	Ground truth Spk emb	Vision-guided Spk emb	ESTOI \uparrow	MCD \downarrow	LSE-C \uparrow	LSE-D \downarrow	WER \downarrow	SECS \uparrow
✓	✗	✗	0.276	9.72	7.16	7.33	40.7%	0.608
✓	✓	✗	0.327	7.85	7.23	7.27	38.4%	0.770
✓	✗	✓	0.284	9.35	7.28	7.27	39.2%	0.625

Table 4. Ablation study on LRS3 dataset analyzing the effectiveness of the vision-guided speaker embeddings. \uparrow means that the higher is the better, and \downarrow means that the lower is the better.

vision-guided speaker extractor actually contains the correct speaker representations in comparing the ground truth ones. To do so, we calculate the cosine similarity between the extracted speaker representations from the generated mel-spectrogram and the ground truth one, respectively, which is called Speaker Encoder Cosine Similarity (SECS). Table 3 shows that the DiffV2S achieves 0.581 SECS and 0.625 SECS on LRS2 and LRS3, respectively, outperforming all the previous methods, even the ones with the speaker embedding with the original audio. This clearly proves that the proposed vision-guided speaker embedding well extracts the speaker representations from the input videos. We also discover that the performances of the previous work without the actual audio input become degraded compared to those of the original ones. This shows that clearly the speaker embeddings extracted from the audio make the output speech more similar to the voice of the actual speech. Most importantly, the DiffV2S can extract the speaker embeddings by utilizing the input video with no need of extra audio information.

5.2. Qualitative Results

We visualize the mel-spectrogram samples generated from the previous works [22, 23, 33] and the proposed method, along with the actual ground truth mel-spectrograms. Figure 4 shows two samples from each dataset, LRS2 and LRS3. It is clearly shown that the samples generated from SVTS [33] and Multi-task [23] tend to be blurry and fail to bring out the details of the actual mel-spectrogram. This kind of tendency would eventually affect the waveform generation, producing noisy speech. The results from VCA-GAN [22] seem to produce the details of the mel-spectrogram; however, compared to the ground truth mels, they fail to fully follow the actual mel-spectrogram, thus producing the wrong audio speech in the end. In contrast, the generated mel-spectrogram samples from the proposed DiffV2S not only adequately represent the fine details of the mel frequency but also visually well match the ground truth mel-spectrograms. These proper and detailed mel-sepectrograms eventually lead to the speech containing the right contents without any noise. The demo video and the audio samples of the generated speech are available in the GitHub repository¹.

¹<https://github.com/joannahong/DiffV2S>

5.3. Ablation Study

We conduct the ablation study to analyze how our vision-guided speaker embeddings extracted from the input video are effective compared to the actual speaker embeddings extracted from the ground truth audio waveforms, shown in Table 4. The baseline refers to the visual encoder with the pre-trained model along with the vanilla diffusion model conditioned on the visual features f_v only. The ground truth speaker embedding is the speaker embedding guidance s_G mentioned in Section 3.1.2, which are directly extracted from the audio speech through the pre-trained speaker encoder. Lastly, the vision guided speaker embedding is the speaker representations directly extracted from the input video.

The results verify that the vision-guided speaker embeddings actually help for the overall quantitative performance, showing the performance improvement in all metrics compared to the baseline architecture. Further, the scores from the vision-guided speaker embeddings show the comparable performance with those from the actual speaker embeddings, even outperforming on several metrics, LSE-C and WER. We can infer from the results that the generated speech is affected from the vision-guided speaker embedding to contain the speaker’s characteristics such as accents, so that it becomes more similar to the actual speech.

6. Conclusion

We propose a novel vision-guided speaker embedding extractor using the pre-trained model and prompt tuning technique. To do so, the rich speaker embedding information is produced from solely on input visual information so that extra audio information is not necessary during the inference. Using the extracted vision-guided speaker embedding representations, we propose the DiffV2S, the conditional diffusion model conditioned on the visual information and the vision-guided speaker embeddings. The DiffV2S not only saves phoneme details contained in the input video frames, but also creates a highly intelligible mel-spectrogram in which the speaker identities of the multiple speakers are all maintained.

Acknowledgements This work was partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2C2005529).

References

- [1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. **2, 5**
- [2] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. Lip2audspec: Speech reconstruction from silent lip movements video. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2516–2520. IEEE, 2018. **2**
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. **2**
- [4] Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Maria Aluisio, and Moacir Antonelli Ponti. Sc-glowtts: An efficient zero-shot multi-speaker text-to-speech model. In *Proc. Interspeech 2021*, pages 3645–3649, 2021. **6**
- [5] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021. **2**
- [6] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. Wavegrad 2: Iterative refinement for text-to-speech synthesis. In *Proc. Interspeech 2021*, pages 3765–3769, 2021. **2**
- [7] Jeongsoo Choi, Minsu Kim, and Yong Man Ro. Intelligible lip-to-speech synthesis with speech units. *arXiv preprint arXiv:2305.19603*, 2023. **1, 2**
- [8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Proc. Interspeech 2018*, pages 1086–1090, 2018. **4**
- [9] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE, 2017. **2, 5**
- [10] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision—ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. **6**
- [11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. **5**
- [12] Ariel Ephrat and Shmuel Peleg. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2017. **2**
- [13] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. **6**
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. **2, 4**
- [15] Joanna Hong, Minsu Kim, Se Jin Park, and Yong Man Ro. Speech reconstruction with reminiscent sound via visual voice memory. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3654–3667, 2021. **1, 2**
- [16] Joanna Hong, Minsu Kim, and Yong Man Ro. Visagesyntalk: Unseen speaker video-to-speech synthesis via speech-visage feature selection. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 452–468. Springer, 2022. **2**
- [17] Jesper Jensen and Cees H Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016. **6**
- [18] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-TTS: A Denoising Diffusion Model for Text-to-Speech. In *Proc. Interspeech 2021*, pages 3605–3609, 2021. **1, 3**
- [19] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022. **2**
- [20] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *arXiv preprint arXiv:2212.13344*, 2022. **5**
- [21] Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. Multi-modality associative bridging through memory: Speech sound recollected from face video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 296–306, 2021. **1**
- [22] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip to speech synthesis with visual context attentional gan. *Advances in Neural Information Processing Systems*, 34, 2021. **1, 2, 6, 7, 8**
- [23] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip-to-speech synthesis in the wild with multi-task learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. **1, 2, 6, 7, 8**
- [24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. **6**
- [25] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020. **6**
- [26] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. **2**
- [27] Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, pages 125–128. IEEE, 1993. **6**
- [28] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. **1, 2, 3**

- [29] Songxiang Liu, Dan Su, and Dong Yu. Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans. *arXiv preprint arXiv:2201.11972*, 2022. 3
- [30] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 2
- [31] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021. 2
- [32] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 1
- [33] Rodrigo Mira, Alexandros Haliassos, Stavros Petridis, Björn W Schuller, and Maja Pantic. Svts: Scalable video-to-speech synthesis. In *Proc. Interspeech 2022*, pages 1836–1840, 2022. 1, 2, 5, 6, 7, 8
- [34] Rodrigo Mira, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, Björn W Schuller, and Maja Pantic. End-to-end video-to-speech synthesis using generative adversarial networks. *IEEE Transactions on Cybernetics*, 2022. 1, 2
- [35] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Proc. Interspeech 2017*, pages 2616–2620, 2017. 4
- [36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [37] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 4
- [38] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021. 3
- [39] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805, 2020. 1, 2
- [40] Bowen Shi, Wei-Ning Hsu, Kushal Lakhota, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022. 1, 5
- [41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2
- [42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 5
- [43] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1
- [44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 6
- [45] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Video-driven speech reconstruction using generative adversarial networks. In *Proc. Interspeech 2019*, pages 4125–4129, 2019. 2
- [46] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE, 2018. 4, 6
- [47] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, 2021. 2