Generation of Facial Images Reflecting Speaker Attributes and Emotions Based on Voice Input

Kotaro Koseki, Yuichi Sei[a], Yasuyuki Tahara[b] and Akihiko Ohsuga[c]
The University of Electro-Communications, Tokyo, Japan

Keywords: Multimodal Learning, Deep Learning Machine Learning, CNN, VAE.

Abstract: The task of "face generation from voice" will bring about a significant change in the way voice calls are made. Voice calls create a psychological gap compared to face to face communication because the other party's face is not visible. Generating a face from voice can alleviate this psychological gap and contribute to more efficient communication. Multimodal learning is a machine learning method that uses different data (e.g., voice and face images) and is being studied to combine various types of information such as text, images, and voice, as in google's imagen(Saharia et al., 2022). In this study, we perform multimodal learning of speech and face images using a CNN convolutional speech encoder and a face image variational autoencoder (VAE: Variational Autoencoder) to create models that can represent speech and face images of different modalities in the same latent space. Focusing on the emotional information of speech, we also built a model that can generate face images that reflect the speaker's emotions and attributes in response to input speech. As a result, we were able to generate face images that reflect rough emotions and attributes, although there are variations in the emotions depending on the type of emotion.

1 INTRODUCTION

The telephone experience will greatly evolve when it becomes possible to recognize a person's face and facial expressions through voice alone. Currently, video calling tools such as ZOOM and TEAMS are widely used in business and academic settings in the wake of the coronavirus. Thus, video calls are more informative than voice calls and are effective for quickly and accurately ascertaining the other party's intentions. On the other hand, however, video calls are not used as much as voice calls in everyday situations, such as a quick phone call with an acquaintance or a sales call.

One of the reasons why video calls are not routinely used is that senders find it inconvenient or uncomfortable to have their faces reflected in the camera. To address these obstacles to the use of video calls, this study proposes a model that generates face images that reflect the speaker's emotions from audio information. This model uses a speech encoder and a Variational AutoEncoder (VAE)(Kingma and Welling, 2013) to learn the relationship between speech and the speaker's face image, and generates a face image from speech. Specifically, the VAE is first trained with the face image as input and output, and the speech encoder section is trained so that the features obtained when the speaker's voice is input to the speech encoder are similar to the features obtained from the VAE encoder section when the speaker's face image is input to the VAE. By replacing the voice encoder trained in this way with the encoder part of the VAE, a face image is generated using the voice as input. In addition, emotion label information was added to the features obtained from the face images by the encoder during VAE training in order to reflect emotion information more strongly in the generated images. This made it possible to manipulate the facial expressions in the VAE reconstructed images according to the emotion information and generate facial images that more closely resemble the facial expressions of the speaker during speech. The task of generating facial images and facial expressions of a speaker from speech information has already been realized using CNN-based models(Zhou et al., 2020)(Oh et al., 2019)(Ji et al., 2021). However, existing research results do not include a method to generate a face image that includes the speaker's attributes and facial ex-

[a] https://orcid.org/0000-0002-2552-6717
[b] https://orcid.org/0000-0002-1939-4455
[c] https://orcid.org/0000-0001-6717-7028

Koseki, K., Sei, Y., Tahara, Y. and Ohsuga, A.
Generation of Facial Images Reflecting Speaker Attributes and Emotions Based on Voice Input.
DOI: 10.5220/0011630200003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 2*, pages 99-105
ISBN: 978-989-758-623-1; ISSN: 2184-433X

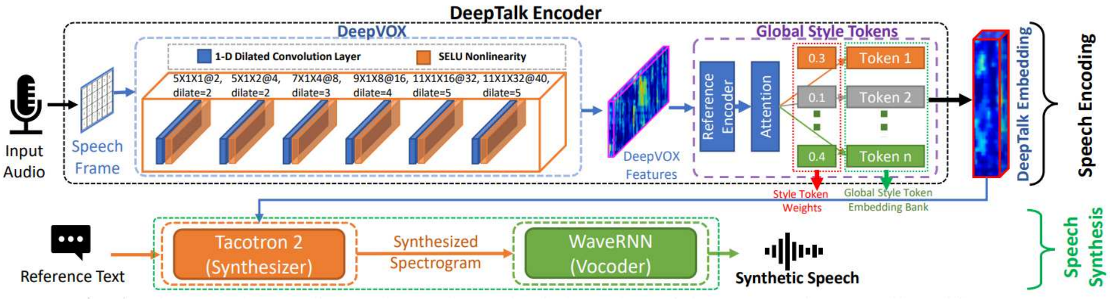

Figure 1: DeepTalk Model Quote from (Chowdhury et al., 2021).

pressions from speech alone, although some of them generate an average face image of the speaker or generate facial expressions by predicting landmarks of facial parts without generating the image itself. Therefore, the results of this research will enable the realization of a simulated video call experience in which the receiver can unilaterally check the facial image including the other party's facial expressions only by his/her own will, even if the other party does not activate the camera, thereby improving the convenience of video calls and the quality of the telephone experience.

This paper is organized as follows. Thereafter, Section 2 of this paper describes related research, Section 3 describes the proposed method, Section 4 describes the experiments, Section 5 discusses the results, and finally Section 6 summarizes the conclusions of this paper and describes future prospects.

2 RELATED WORK

2.1 DeepTalk

DeepTalk(Chowdhury et al., 2021) is a deep learning speaker recognition and speech synthesis model presented by Anurag Chowdhury et al. The speech input to the model is encoded by a CNN and Multi-Head Attention to obtain speaker-specific speech features such as speaker pitch, voice color, and speaking habits. The speech features extracted by the encoder successfully improve the robustness and accuracy of the speaker recognition task. They are also used for tasks such as generating synthesized speech using the speech features. A schematic diagram of the model is shown in Figure 1.

2.2 Variational AutoEncoder

Variational AutoEncoder (VAE)(Kingma and Welling, 2013) is a model that incorporates the idea of using random variables calculated from the distribution of the latent space as an intermediate representation of AutoEncoder. While general AutoEncoders often use convolutionally processed features of input data as latent variables, VAE calculates mean μ and variance σ^2 for convolutionally processed features, and latent variable z is obtained from mean μ and variance σ^2. Since the mean μ and variance σ^2 are parameters of the distribution of the latent space z, sampling is necessary to obtain the latent variable z in VAE. However, sampling would prevent error back propagation. Therefore, in VAE, the latent variable z is defined by the following equation, assuming that ε follows a normal distribution.

$$z = \mu + \varepsilon\sigma \tag{1}$$

$$\varepsilon \sim \mathcal{N}(0,1) \tag{2}$$

This makes it possible to pseudo-sample the latent variable z from its distribution in the latent space and to perform error back propagation because the latent variable z is differentiable. In general, the following equation is used for the loss function.

$$\mathcal{L}_{\mathrm{VAE}}[\mathrm{q}_\varphi(\mathrm{z}|\mathrm{X})] = \mathrm{E}_{\mathrm{q}_\varphi}[\log \mathrm{p}_\theta(\mathrm{X}|\mathrm{z})] - \mathrm{KL}[\mathrm{q}_\varphi(\mathrm{z}|\mathrm{X})\|\mathrm{p}(\mathrm{z})] \tag{3}$$

In Equation 3, the first term on the right-hand side is the term used to maximize the log-likelihood of returning the latent variable z to the input image X. In this study, it was replaced by the mean-square error to account for semantic similarity. The second term represents the KL divergence distance from the prior distribution to the posterior distribution.

3 METHOD

3.1 Overview

First, a schematic diagram of the model for generating face images from input speech using VAE and a

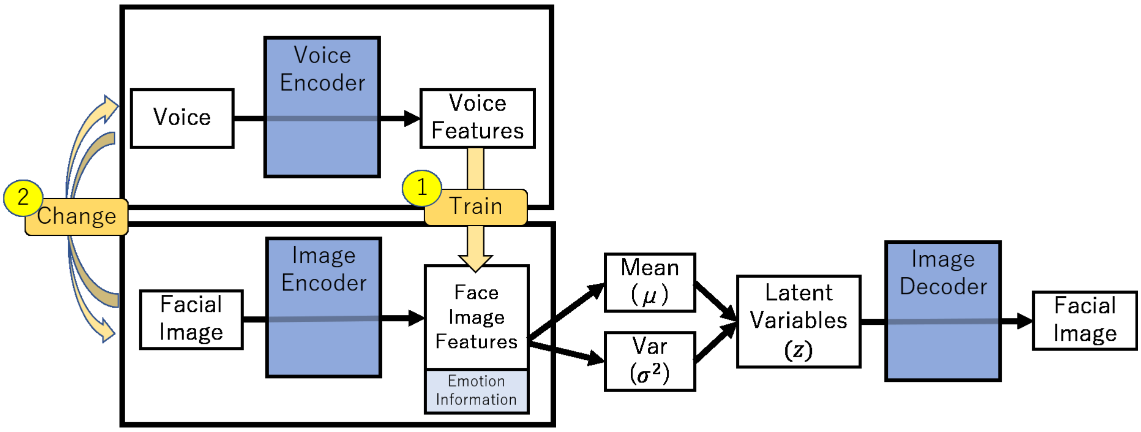

Figure 2: Overall diagram of the proposed method.

voice encoder is shown in Figure 2. In this study, we first create a VAE model that can reconstruct the same face image as the input face image. Next, the voice encoder is trained to be able to extract the same feature values for the face features output by the encoder of the VAE model when the voice corresponding to the face image is input. This means that the voice features obtained when a speaker's voice is input to the voice encoder are similar to the facial features obtained when the speaker's face image is input to the encoder of the VAE model. Therefore, by calculating the latent variable z from the output of the voice encoder and decoding it using the decoder of the VAE model, a face image of the speaker corresponding to the input voice can be generated. The following sections 3.2 and 3.3 describe the structure of the VAE and voice encoder models used and the advantages of using each model, while section 3.4 describes the labeling process used to strongly express the emotional information in the reconstructed images.

3.2 VAE

VAE and Generative Adversarial Network (GAN)(Goodfellow et al., 2020) are the main models that output images in response to input. In this study, we used VAE among image generation models, focusing on the property of AutoEncoder (AE)(Hinton and Salakhutdinov, 2006) that it can reconstruct the input image. Unlike ordinary AE, latent variables in VAE are based on a probability distribution and are learned using the following equation 3 generally. Therefore, compared to AE, the latent space to be learned is continuous. Since it is impossible to learn all human voices and faces, it is not always possible to perfectly reproduce a speaker's face in response to input speech, but we used VAE because we expected that this feature of VAE would allow us to generate a face that is comfortable in response to input speech.

3.3 Voice Encoder

The DeepTalk model described in 2.1 shows excellent accuracy for the speaker recognition task. In the speaker recognition task, it is more important to learn tonal features such as pitch and accentuation well than the content of each speaker's speech. In this study, the encoder part of the DeepTalk model is used as a speech encoder because learning tonal features is more important than the content of speech. The specific model structure is shown in Table 1. In this model, the audio waveform is first convolved, then the temporal waveform is convolved, and finally multiple attachments are computed and combined by Multi-HeadAttention to obtain features of the same dimension as those extracted by the encoder of the VAE model. By learning between the audio features obtained by this operation and the face image features output from the VAE encoder, it is possible to learn to match the output from the VAE encoder of the corresponding face image with the output from the audio encoder.

3.4 Emotion Label Addition

The purpose of this study is to generate face images with facial expressions that are not expressionless, but rather in accordance with the emotional information in the input audio. Therefore, when training VAE, we added the emotional information of the input image to the convolutional features before finding the latent variable z. Four emotion labels, Angry, Sad, Happy, and Neutral, were added to the convolutional features as 8-dimensional information as shown below.

- Angry: 00000011
- Sad: 00001100
- Happy: 00110000
- Neutral:11000000

Although the four types of information can be added with a minimum of two digits, eight digits of information were added in this study. The reason for this is that 2-digit information has little effect on the reconstructed image when emotional information is added, and through experiments, we confirmed that 8-digit information is optimal considering the trade-off between the quality of the reconstructed image and the effect of the emotional label.

Table 1: VoiceEncoder Model.

Input: (batchsize(32) × 1 × 160 × 200)				
type	kernel size	stride	dilation	output shape
Conv2d	5 × 1	1 × 1	2 × 1	32×2×152×200
SELU				32×2×152×200
Conv2d	5 × 1	1 × 1	2 × 1	32×4×144×200
SELU				32×4×144×200
Conv2d	7 × 1	1 × 1	3 × 1	32×8×126×200
SELU				32×8×126×200
Conv2d	9 × 1	1 × 1	4 × 1	32×16×94×200
SELU				32×16×94×200
Conv2d	11 × 1	1 × 1	5 × 1	32×32×44×200
SELU				32×32×44×200
ZeroPad2d				32×32×51×200
Conv2d	11 × 1	1 × 1	5 × 1	32×40×1×200
SELU				32×40×1×200
type	kernel size	stride	padding	output shape
Conv2d	3 × 3	2 × 2	1 × 1	32×32×100×20
BatchNorm2d				32×32×100×20
Conv2d	3 × 3	2 × 2	1 × 1	32×32×50×10
BatchNorm2d				32×32×50×10
Conv2d	3 × 3	2 × 2	1 × 1	32×64×25×5
BatchNorm2d				32×64×25×5
Conv2d	3 × 3	2 × 2	1 × 1	32×64×13×3
BatchNorm2d				32×64×13×3
Conv2d	3 × 3	2 × 2	1 × 1	32×128×7×2
BatchNorm2d				32×128×7×2
Conv2d	3 × 3	2 × 2	1 × 1	32×128×4×1
BatchNorm2d				32×128×4×1
GRU				32×4×200
MultiHeadAttention				32×1×256
Output: (*batchsize(32)* × 1 × 256				

4 EXPERIMENTS

4.1 Dataset

For training data, a dataset consisting of human face images and corresponding speech utterances is required. Therefore, we used the Crema-D(Crowd-sourced Emotional Multimodal Actors Dataset)(Cao et al., 2014), which is an emotion-labeled actor's speech video dataset, composed as follows.

- Number of Speakers : 91
- Male:Female = 48:43
- race : African American, Asian, Caucasian, Hispanic
- Age : 20∼74
- Number of clips : 7442

The following preprocessing was performed on the video and audio data, since the data required for this project is not video and audio but a data set of image and audio. The training and test data were split at a ratio of 8:2 and used for training.

4.1.1 Video Preprocessing

One face image of the speaker is used for each speech video. Therefore, it is necessary to select a specific frame from the speech video as the face image of the speaker. Crema-D is a dataset with emotional information. Therefore, the proposed method selects the frame in the video that most strongly expresses the emotion assigned as a label, and adopts it as the speaker's face image. The python library dlib(King, 2009) was used to crop each face image so that only the face area is within the contour, and grayscale transformation, resizing, and normalization were performed.

4.1.2 Voice Preprocessing

The sampling frequency of the voice files was set to 8000 Hz. In addition, the proposed method removes silent intervals of one second or longer from the speech data, since the purpose of this study is to learn features such as voice tone, not speech content. In order to match the length of the audio data, the data length of all data was iterated to be 3.67 seconds, matching the length of the longest data among the data with the deleted silent segments.

4.2 VAE Training and Generation Experiment

Based on the proposed method, VAEs were trained in the following settings.

- Epoch:300
- Learning Rate:0.001
- Input Size: 64*64 pixel
- Optimizer:Adam
- Batch Size:32

The training curve during training and the face images reconstructed by inputting face images for the trained VAE are shown in Figures 3 and 4, respectively.

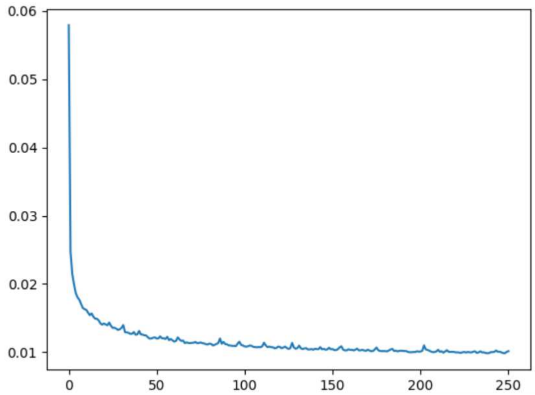

Figure 3: VAE Train Loss.

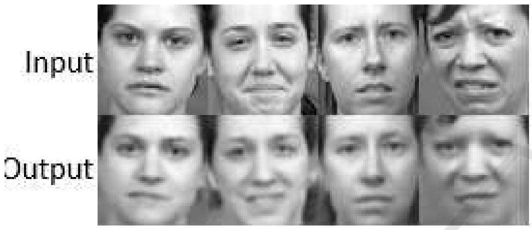

Figure 4: Generated Results by VAE.

4.3 Voice Encoder Training and Generation Experiment

The Voice Encoder was trained in the following settings.

- Epoch:600
- Learning Rate:0.0001
- Loss Function:MSE
- Optimizer:Adam
- Batch Size:32

Figure 6 and Figure 7 show the learning curve during training and the results of generating face images by connecting the trained voice encoder to the VAE as shown in Figure 3, respectively.

5 DISCUSSION

5.1 VAE Training and Generation Experiment

Figures 3 and 4 show that VAE learning is successful without any problems. In particular, Figure 4 shows that the output image faithfully reproduces the attributes of the input image. In addition, as can be seen from the fact that the second row of Figure 5 shows an open-mouth smile instead of a closed-mouth smile, it can be confirmed that the emotional information itself is conveyed even if the facial expressions in the reconstructed images are not exactly the same as those in the input images.

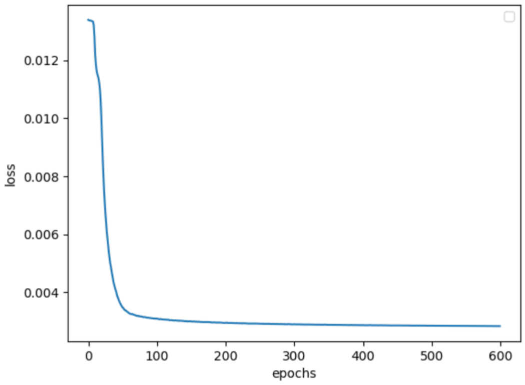

Figure 5: Voice Encoder Train Loss.

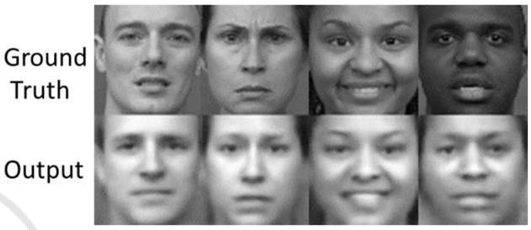

Figure 6: Generated Results by Voice Encoder.

5.2 Effects of Emotion Label Addition

As shown in Section 3.4, in this study, emotion labels are assigned to VAEs when they are learned. From 4.2, we have confirmed that VAE is able to learn emotional information as well, but in this section, we discuss the effect of adding emotional labels on the generated results.

Figure 7 shows the results of reconstructing images by assigning “Angry”, “Sad”, “Happy”, and “Neutral” labels to the convolutional features of the input face images, ignoring the original emotion labels in the input face images. The first row shows the input image for each row, and the second and subsequent rows show the reconstructed images with the respective emotion labels. Figure 8 shows that the “happy” image is successfully converted for all input images, even when the input image clearly expresses negative emotions, since it is possible to generate an image with raised corners of the mouth. On the other hand, the other three emotions showed differences in facial expressions depending on the emotion labels, and although the negative-positive level produced reasonable facial images for all emotion labels, the dif-

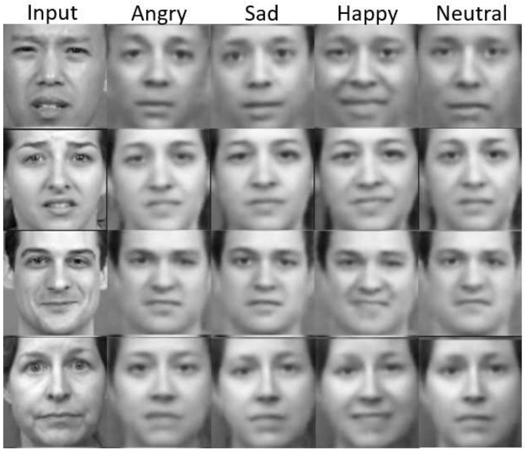

Figure 7: Result of replacing labels.

ferences were not as clear as for the "happy".

The reason for this is assumed to be that there is a wide range of emotions in "angry" and "sad". For example, there are many types of facial expressions that are expressed depending on the type of angery, such as "cross," "annoyed," and "irritated," and it is possible that sad and neutral faces were included in the angery-labeled images, which may have negatively affected learning. On the contrary, "happy" is the only positive emotion, and although there are different types such as "smile," "laugh," and "grin," it has the common feature of raising the corners of the mouth, and we assume that it was learned successfully with low learning difficulty.

5.3 Voice Encoder Training and Generation Experiment

First, from Figure 6, it can be confirmed that the voice encoder was successfully trained with the ImageEncoder features of VAE. Next, the trained voice encoder is integrated into VAE as shown in Figure 3, and the results of face image generation from voice are confirmed. Figure 7 shows that the general characteristics of the face, such as gender, were captured. Although there are some differences in details such as the size of the eyes and the depth of the moat, the accuracy is sufficient for one of the purposes of this study, which is "to understand the face of an unknown person from voice alone". As in 5.2, significant results were obtained for the facial expression "happiness", but for the other emotions, the accuracy was only as good as the identification of positive and negative emotions. We assume that this is also due to the fact that a wide range of emotions are included, as was the case in the VAE training.

6 CONCLUSION

In this paper, we proposed a method for generating face images that reflect the attributes and emotions of the speaker based on voice input for the purpose of conducting a simulated video call using only voice. As a result, the proposed method was not able to learn facial expressions well when negative emotions such as "angry" and "sad" had similar facial expressions, but it was able to reflect emotional information well for emotions with monotonous facial expression features such as "happy". In addition, we were able to generate a face image from the input speech that reflected the attributes of the speaker's face image.

7 FUTURE WORK

This paper does not quantitatively evaluate the generation results using specific indicators. Therefore, it is necessary to quantitatively evaluate the generation results of this model using specific numerical values in the future. In addition, since the purpose of this research is to understand the face and facial expression of an unknown speaker during telephone communication, it is more important to generate the most plausible face image for the input voice than to generate the exact face of the speaker. Therefore, in addition to quantitative evaluation, qualitative evaluation is also an important evaluation item and should be conducted. As a future challenge, we would like to expand the training data to a larger dataset to improve the robustness of the model to all types of input speech, since it cannot be said that we have acquired sufficient speaker diversity with the dataset used in this study.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP21H03496, JP22K12157.

REFERENCES

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Chowdhury, A., Ross, A., and David, P. (2021). Deeptalk: Vocal style encoding for speaker recognition and speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6189–6193. IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C. C., Cao, X., and Xu, F. (2021). Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Oh, T.-H., Dekel, T., Kim, C., Mosseri, I., Freeman, W. T., Rubinstein, M., and Matusik, W. (2019). Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7539–7548.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., and Li, D. (2020). Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15.