Disentangled Audio-Driven NeRF: Talking Head Generation with Detailed Identity-Specific Microexpressions

Seoyoung Lee* Seongsu Ha* Joonseok Lee
Seoul National University
{seoyoung1215, sha17, joonseok}@snu.ac.kr

Abstract

Recent research has enabled the rendering of talking head videos that capture large dynamics of the head with high fidelity. However, modeling the detailed identity-specific microexpressions and spontaneous movements, such as lip movement and eye blinking, while achieving high synchronization between the auditory and visual signals remains a challenge. In this paper, we address this issue with the help of neural implicit functions conditioned on disentangled audio. Specifically, we first extract audio features that are disentangled into core auditory components (content, timbre, rhythm, and pitch) that retain identity-specific information. Then, the disentangled audio embeddings are fed into a conditional implicit function together with visual embeddings so that high quality audio-visual mappings for details are learned. Experimental results demonstrate that our method can (1) successfully render detailed identity-specific microexpressions that are personalized for each person being modeled, and (2) improve fidelity of the audio-visual rendered results in general.

1. Introduction & Related Works

Along with the emergence of deep learning methods for content generation, talking head generation has attracted significant attention due to its abundant applications. Talking head generation aims at synthesizing a realistic target face, which talks in correspondence to the given audio sequences. This task generally requires attention to two aspects. Firstly in terms of dynamics, it is important to capture large dynamics of the head and microexpressions, such as lip movements and eye blinks. Secondly, the generated results should have high synchronization between the two major modalities, auditory and visual. This is not simple as the two modalities are quite different, entailing a difficulty in bridging the inherently large domain gap between the audio and visual signals in generating a talking head.

As the aforementioned issues are not trivial, previous works have shown various approaches, such as 2D [4,12,18] (*e.g.*, landmarks [17]) or 3D [8, 14, 15] intermediate representations (*e.g.*, 3D face models) to model the head. While intermediate representations have advantages, they usually lead to a mismatch in audio-visual mapping due to information loss from mapping to a predefined finite solution set. Recently, models that do not use intermediate representations, but rather direct implicit representations [7, 10, 11], have captured the large dynamics of the talking head with high fidelity, and are suitable for advanced talking head editing tasks [3,5,15,16]. By learning a head model solely from the given video data, these models retain as much information from the overall ground truth as possible.

However, these models are still limited in modeling personalized or identity-specific expressivity via microexpressions (*e.g.*, lip movements), as well as spontaneous movements (*e.g.*, eye blinks and sudden head movements). While the models synthesize general talking behavior well, personalized behaviors, such as different habits in lip movement even when the same word is pronounced, and instantaneous spontaneous movements in only a minority of the frames are not well modeled. Thus, some models create people who do not blink at all or blink unnaturally, as if their eyes are fluttering while half-open, with low fidelity to the ground truth. These aspects can lead to undesirable phenomena, like the uncanny valley [6].

One potential reason for the unnatural modeling in previous models of customized microexpressions and spontaneous movements may be due to a lack of consideration of appropriate audio embeddings for talking head generation. Speech can be represented in various components, such as content, pitch, timbre, and rhythm. However, speech models (e.g., DeepSpeech [1]) used in previous models build audio features so that audio signals are recognized altogether without consideration of the different components of speech. We propose that disentangling multiple styles or prosody information in audio can help with more expressive and controlled speech and talking face synthesis.

Based on these unsolved problems in audio-driven talking head generation, our contributions are as follows:

[*] These authors contributed equally.

- We present a model that can learn detailed identity-specific microexpressions and spontaneous movements, such as lip movement and eye blinking, with disentangled audio features.
- Our disentangled audio representations allow audio-visual mapping with higher modality correspondence.

2. Preliminaries

NeRF [10] models 3D implicit representations of static objects or scenes via an MLP, which maps a 3D location $\mathbf{x} = (x, y, z)$ and viewing direction $\mathbf{d} = (\theta, \phi)$ to the corresponding volumetric density σ and emitted color $\mathbf{c} = (r, g, b)$. The color of pixel $C(r)$ along a camera ray r is estimated by accumulating the color and transmittance of N sampled points along the ray, based on density σ:

$$\hat{C}(r) = \sum_{i=0}^{N-1} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) \mathbf{c}_i, \text{ where } T_i = \exp\left(-\sum_{j=0}^{i-1} \sigma_j \delta_j\right), \tag{1}$$

where δ_i is the distance between adjacent sampled points. NeRF is optimized by comparing estimated colors $\hat{C}(r)$ for a batch of rays $\mathcal{R}$ and their ground truth pixel colors:

$$\mathcal{L} = \sum_{r \in \mathcal{R}} \left[\|C(r) - \hat{C}(r)\|_2^2 \right] \tag{2}$$

AD-NeRF [5] presents a conditional NeRF with a semantic feature of audio $\mathbf{a}$ as an additional input, along with viewing direction $\mathbf{d}$ and 3D location $\mathbf{x}$. In training NeRF for the head, rigid face pose parameters $\Pi = \{R, t\}$ are used to transform sampling points of the head to the canonical space. Then a NeRF for the torso is trained sequentially in a similar manner, except with Π as an additional input:

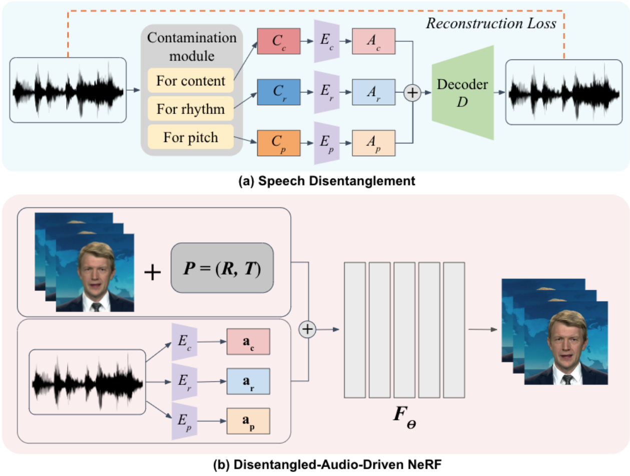
$$F_\theta^{head} : (\mathbf{a}, \mathbf{d}, \mathbf{x}) \rightarrow (\mathbf{c}, \sigma) \tag{3}$$

$$F_\theta^{torso} : (\mathbf{a}, \mathbf{d}, \mathbf{x}, \Pi) \rightarrow (\mathbf{c}, \sigma) \tag{4}$$

For audio $\mathbf{a}$, semantically meaningful information was extracted from acoustic signals via the DeepSpeech [1] model. DeepSpeech is an end-to-end speech recognition, or speech-to-text model known to be robust to noise, echo, and various speaker-specific properties like pitch.

3. Method

Our model takes disentangled audio features to perform audio-driven talking head generation that achieves to model detailed identity-specific microexpressions using neural radiance fields. To attain disentangled content, timbre, rhythm, and pitch from one audio source, we reference SpeechSplit [13]. The disentangled audio embeddings obtained from the encoders of SpeechSplit are then used in combination with visual information in the form of 3D location and viewing direction, to get the colors and densities for each 3D location with a certain audio via NeRF.



(a) Speech Disentanglement

(b) Disentangled-Audio-Driven NeRF

Figure 1. **An overview of our model.** (a) Audio is disentangled into content, rhythm, and pitch for each speaker. (b) The disentangled auditory components are concatenated with visual features as input to NeRF.

In Sec. 3.1, we first present how an audio is disentangled. Next, we describe how we achieve fine-grained head rendering via NeRF with the decomposed audio embeddings.

3.1. Identity-specific Audio Representation

Human speech can be roughly decomposed into four important components: content, timbre, pitch and rhythm. Our approach shown in Fig. 1(a) is built upon SpeechSplit [13] to disentangle an audio to achieve sophisticated and discriminative audio representations for rendering a more natural-looking talking person. SpeechSplit has an encoder-decoder architecture trained to reconstruct the given mel-spectrogram, S. The encoder contains three parallel sub-encoders E_c, E_r, E_p that take the outputs of the contamination module. Different audio components (content, rhythm, and pitch) are contaminated respectively as C_c, C_r, C_p to limit information each sub-encoder can see. We exclude timbre in this step, because our model is trained to learn a video of one person, so there already is only one voice. Since every input to each sub-encoder is partially contaminated, each sub-encoder works as an information bottleneck trained to extract one of the intact audio components, A_c, A_r or A_p, to provide all necessary information to reconstruct the original audio mel-spectrogram.

$$A_c = E_c(C_c), \quad A_r = E_r(C_r), \quad A_p = E_p(C_p) \tag{5}$$

The decoder D reconstructs a version of the original audio mel-spectrogram $\hat{S}$ by combining these disentangled audio embeddings from the sub-encoders.

$$\hat{S} = D(A_c, A_r, A_p) \tag{6}$$

3.2. Audio-visual Mapping for Details

After training the encoders for audio disentanglement in Sec. 3.1, we extract the disentangled content, rhythm, and pitch from the respective encoders. Then as in AD-NeRF, but with better disentangled auditory features, content $\mathbf{a_c}$, rhythm $\mathbf{a_r}$ and pitch $\mathbf{a_p}$, we concatenate these with $\mathbf{x}$ and $\mathbf{d}$ as an input to train a single NeRF to get the colors and densities of each 3D location when a certain audio clip is accompanied. As shown in Fig. 1(b), the neural function can be formulated as:

$$F_\theta : (\mathbf{a_c}, \mathbf{a_r}, \mathbf{a_p}, \mathbf{d}, \mathbf{x}) \rightarrow (\mathbf{c}, \sigma) \qquad (7)$$

The general procedure on training and volume rendering of the NeRFs are same as those in AD-NeRF [5].

4. Experiments

4.1. Implementation Details and Computation Time

The original repository of AD-NeRF suffers from a long training time due to data bottleneck with no support of batching and multi-GPU training. In order to facilitate our research process, we implemented our framework from scratch in PyTorch-Lightning and Hydra. Thanks to the new dataloader and framework, both training and inference time were reduced by less than 1/2 on a single GPU and 1/5 with four GPUs with a batch size of 4. All experiments were run on V100 GPUs and the same hyper-parameters were used as in SpeechSplit and AD-NeRF.

4.2. Evaluation metrics

We measure PSNR and SyncNet [2] scores to prove the superiority of our method. SyncNet takes a pair of input sequences to make a correlation based deep neural network with anti-causal convolution sets and can be used to estimate time delay for two different input sources. We use SyncNet scores to measure synchronization quality between audio and visual signals in talking head generation.

4.3. Overall Results

Quantitative Anaylsis Quantitative results are reported in Tab. 1. Our method, using content, rhythm, and pitch embeddings from SpeechSplit performs better than our baseline using DeepSpeech features as the audio embedding in both PSNR and SyncNet scores.

Qualitative Analysis Results in the quantitative analysis are backed up with qualitative results in the rendered lips and eyes. Based on the generated video, our model increases the lip detail and renders more natural eye blinks, so that no more blurry or fluttering eye blinks occur. Pixel difference heat map between the ground truth and rendered images in Fig. 2 clearly show that our approach is better at capturing not only the overall face, but also the eyes and fine details of the face. The general fidelity to the overall face and eye blinks seems to be in line with the increase in PSNR, while the fidelity to the lip movements tend to be proportional to the increasing SyncNet scores.

	SpeechSplit Rhythm	DeepSpeech [1, 5]	SpeechSplit Pitch	SpeechSplit Content	SpeechSplit (Ours)	DeepSpeech and SpeechSplit
PSNR(dB)↑	23.48	24.98	25.82	26.77	27.49	27.64
SyncNet score↑	3.426	5.355	5.252	5.478	5.555	5.558

Table 1. Quantitative Results. We report PSNR and SyncNet results for different types of audio embeddings on a video of Obama.

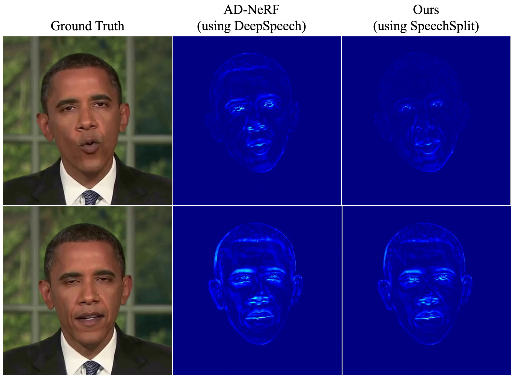


Figure 2. Heat map of pixel differences on different audio embeddings. Rendered at the same frame with just different audio embeddings, our model using SpeechSplit for auditory representations shows more similar pixel values to the ground truth frame.

Discussion We speculate that this is due to the properties of DeepSpeech and SpeechSplit. First, DeepSpeech is trained with a speech-to-text task while SpeechSplit is trained via audio reconstruction. Also, DeepSpeech features are robust to noise and variance in speaker traits, while SpeechSplit disentangles various auditory properties and preserves speaker-specific information. As a result, when we learn the audio-visual mapping with DeepSpeech, detailed lip and eye movement show a general trend, instead of the person-specific result, creating more blurry artifacts.

4.4. Ablation Study

Furthermore, we show ablation studies on various audio embeddings. First, rendering quality improves in order of rhythm, pitch, and content embeddings when used independently as shown in Tab. 1. These quantitative results are in line with qualitative results, approached in a patch-based method [9] that focuses on important areas of interest, which in our case are eyes and lips. When ground truth has two blinks, Fig. 3 shows that using just rhythm only generates one blink with a blurry lower lip. DeepSpeech shows unnatural fidgety blinks, while pitch embeddings show relatively lower audio-visual sync when used solely. Using DeepSpeech and SpeechSplit together also increases PSNR

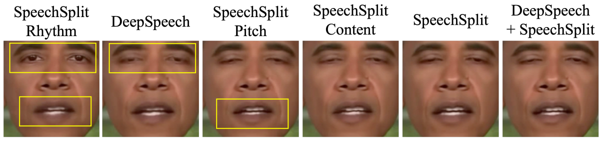


Figure 3. Results with different speech embeddings at blinks. Images are ordered from left to right in order of PSNR.

and SyncNet scores as shown in Tab. 1, implying that DeepSpeech and SpeechSplit may hold different information that are complementary to each other.

5. Conclusion

We have presented a method for high-fidelity talking head synthesis that renders detailed identity-specific microexpressions and spontaneous movements that are personalized for each person. Our method improves the general fidelity and realism of the rendered frames. For future research, we could work on improving the audio-visual embeddings that allow not just accurate renderings, but also more sophisticated audio-visual generation and manipulation. Various methods, such as contrastive learning or self-supervised learning, can be applied to disentangle speech components suitable for learning high-quality audio-visual relationships in different tasks.

References

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016. 1, 2, 3

[2] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 3

[3] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 1

[4] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia*, 2021. 1

[5] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 1, 2, 3

[6] Jiaqi Hao, Shiguang Liu, and Qing Xu. Controlling eye blink for talking face generation via eye conversion. In *SIGGRAPH Asia 2021 Technical Communications*, pages 1–4. 2021. 1

[7] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020. 1

[8] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 1

[9] Sangho Lee, Seoyoung Lee, and Joonseok Lee. Learning to wear: Details-preserved virtual try-on via disentangling clothes and wearer. In *The 33rd British Machine Vision Conference, BMVC 2022*. British Machine Vision Association, 2022. 3

[10] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2

[11] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 1

[12] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 1

[13] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pages 7836–7846. PMLR, 2020. 2

[14] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598, 2022. 1

[15] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*, pages 716–731. Springer, 2020. 1

[16] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 1

[17] Lingyun Yu, Jun Yu, and Qiang Ling. Mining audio, text and visual information for talking face generation. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 787–795. IEEE, 2019. 1

[18] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9299–9306, 2019. 1