LipNeRF: What is the right feature space to lip-sync a NeRF?

Aggelina Chatziagapi[1], ShahRukh Athar[1], Abhinav Jain[2], Rohith MV[2], Vimal Bhat[2], Dimitris Samaras[1]

[1] Stony Brook University [2] Amazon Prime Video

Fig. 1. LipNeRF enables high quality lip-syncing of cinematic HD content, handling challenging expressions, head poses and illumination.

***Abstract*— Synthesizing high-fidelity talking head videos of an arbitrary identity, lip-synced to a target speech segment, is a challenging problem. Recent GAN-based methods succeed by training a model on a large amount of videos, allowing the generator to learn a variety of audio-lip representations. However, they are unable to handle head pose changes. On the other hand, Neural Radiance Fields (NeRFs) model the 3D face geometry more accurately. Current audio-conditioned NeRFs are not as good in lip synchronization as GANs, since they are trained on limited video data of a single identity. In this work, we propose LipNeRF, a lip-syncing NeRF that bridges the gap between the accurate lip synchronization of GAN-based methods and the accurate 3D face modeling of NeRFs. LipNeRF is conditioned on the expression space of a 3DMM, instead of the audio feature space. We experimentally demonstrate that the expression space gives a better representation for the lip shape than the audio feature space. LipNeRF shows a significant improvement in lip-sync quality over the current state-of-the-art, especially in high-definition videos of cinematic content, with challenging pose, illumination and expression variations.**

I. INTRODUCTION

Humans are sensitive to the synchronization of lip motion with speech. In dubbed movies, we always notice the stark misalignment between the movement of the lips, which follow the original audio, and the dubbed audio. Synthesizing photorealistic lip motion, synced with a target audio track, remains a challenging problem. Recent methods for audio-driven talking head video synthesis propose paradigms based on generative adversarial networks (GANs) [32], [42], [51]. They leverage large video datasets with multiple identities, learning a large portion of the audio-lip feature space. However, since these models only act in the 2D image space, they are unable to model the 3D face geometry and handle large head pose variations. In contrast, Neural Radiance Fields (NeRFs) [29] implicitly represent 3D information. Nevertheless, current audio-conditioned NeRFs [19], [45], [27] do not produce as plausible lip synchronization as GAN-based methods, since they are trained on limited video data of a single identity.

In this work, we propose LipNeRF, a novel method that performs lip-syncing in the expression space of a 3D morphable model (3DMM). We empirically demonstrate that the expression space is a far superior feature space for lip-syncing than the audio feature space, using an identity specific training paradigm. During inference, we propose a simple procedure that maps the target audio features to expression parameters, leveraging the lip sync accuracy of a pre-trained GAN-based model. In this way, LipNeRF bridges the gap between the accurate lip synchronization of GAN-based methods and the accurate 3D face modeling of NeRFs.

Additionally, we collected a challenging dataset to evaluate cinematic dubbing. Most related works evaluate their method on datasets in English, using the same audio from the source video (reconstruction) or using a randomly sampled audio from a different speaker. To the best of our knowledge, there is no publicly available dataset that includes talking head videos from movies in HD quality with corresponding dubbed audio in different languages. Compared to current public datasets, our dataset is more challenging due to the expressiveness of the actors, the emphatic head movements and the cinematic lighting. We use this dataset to demonstrate the superiority of LipNeRF over the current state-of-the-art.

In brief, the contributions of our work are as follows:

- We propose LipNeRF, a novel NeRF-based method that performs lip-syncing in the expression space, instead of the audio feature space, achieving high lip sync accuracy for challenging videos of HD quality.
- We collected a dataset of HD talking faces from movie scenes, with corresponding dubbed audio in different languages, appropriate for addressing the problem of lip-syncing for movie dubbing.
- We evaluated quantitatively and qualitatively our

method on the proposed dataset, demonstrating its superiority over the current state-of-the-art.

II. RELATED WORK

Audio-driven Talking Head Video Synthesis. Earlier approaches for lip synced video synthesis, such as Video Rewrite [8] and Voice Puppetry [7], propose probabilistic models that map the phonemes of an audio sequence to corresponding mouth shapes (visemes). This phoneme-to-viseme mapping can be learned by HMMs [34], [16], decision trees [23], or long short-term memory (LSTM) units [15], [36]. However, most of these methods require phoneme labels with millisecond-accurate timestamps, that are usually extracted from error-prone speech-to-text systems. More recent methods avoid the explicit video segmentation into phonemes and visemes. Synthesizing Obama [39] produces photorealistic lip synced videos of President Obama, leveraging a large amount of video footage (17 hours) for training. Speech2Vid [21] proposes an encoder-decoder architecture, where each input face image is conditioned on the corresponding speech segment. MakeItTalk [52] addresses the problem of single image animation, using 3D facial landmarks as an intermediate representation. Neural Voice Puppetry [40] learns an audio-to-expression mapping, based on a 3D face model. Similarly, LipSync3D [25] trains a speaker-specific model, regressing the 3D face geometry and texture for every video frame. Our method also uses the expression space of a 3DMM. However, it is based on NeRFs that are able to model the 3D geometry more accurately and synthesize higher visual quality images.

GAN-based methods are trained on large datasets of videos with multiple identities, learning a large portion of the audio-lip product space. SDA [42] proposes a temporal GAN that animates an input image through a series of RNN layers. ATVG [10] uses 2D facial landmarks as an intermediate representation. DAVS [50] and PC-AVS [51] learn audio-visual representation that disentangles identity-related and speech-related information. PC-AVS [51] enables additional head pose control, using a pose source video as input. Wav2Lip [32] achieves a highly competitive lip sync accuracy. It proposes a convolutional encoder-decoder architecture, conditioned on mel-spectograms and trained to minimize a lip sync expert loss and an L1 reconstruction loss. The main drawback of GAN-based models is that they are strictly 2D and thus unable to handle large pose variations.

In contrast, NeRF [29] based methods represent a scene using a 3D volume where each point is associated with a radiance and density. In this way, they can model the 3D face geometry more accurately. AD-NeRF [19] proposes a dynamic NeRF, conditioned on DeepSpeech [20] audio features. It is trained on a single-identity video of 3-5 minutes length. DFA-NeRF [45] conditions a NeRF on disentangled representations that capture lip motion and personalized face attributes. SSP-NeRF [27] proposes a semantic-aware dynamic ray sampling, based on the intuition that different face regions correlate differently with speech. However, all these NeRF-based approaches, being identity specific, cannot not achieve as good lip synchronization as GAN-based methods. Our method generates high quality lip-syncing results, trained on a single video of even only 20 seconds duration, of cinematic content, with challenging pose, illumination, and expression variations.

Facial Expression Control. Over the years there have been a number of works that perform high quality facial expression editing for 2D images [37], [38], [2], [33], [11], [12]. In GANimation [33] and DefGAN [2], convolutional networks are trained with cycle consistency constraints in an unsupervised manner. Other methods use a 3DMM to reanimate faces [22], [14], [24], [1], [41]. More recently, NeRFs have made it possible to model the 3D face geometry more accurately, leading to more photorealistic results. NerFACE [17], FLAME-in-NeRF [3] and RigNeRF [4] provide facial expression and head pose controls in 3D for high quality reanimation. However, since all these models work in the expression space, they cannot be used for lip-syncing from only audio input, i.e. when the video of a dubbing artist is not available. In contrast, LipNeRF includes a way to map audio to expression, allowing accurate lip synchronization for any target dubbed audio.

III. METHOD

A. Overview

We present LipNeRF, a method that synthesizes high-quality audio-driven talking head videos with accurate lip synchronization. An overview of our approach is shown in Figure 2. Given a talking head video, we first fit a 3DMM and extract the head pose and expression parameters per frame. These parameters condition a dynamic NeRF, in order to generate the corresponding lip-synced talking face. Additionally, we learn per-frame latent codes that intend to capture the appearance of each video frame [30], [17], [4], [28]. Cinematic videos have small per-frame variations, caused by environmental factors, which are independent of audio and expression, and it is necessary to reconstruct them accurately. During inference, we map the target audio to the corresponding expression parameters, leveraging a pre-trained GAN-based model [32]. The extracted expressions, along with the learned latent codes are used as input to LipNeRF. In the following paragraphs, we describe in more detail each individual component of our approach.

B. Dynamic NeRF

LipNeRF learns a conditional NeRF that models dynamic movements of the speaker's face and head. For each source video frame, we segment the head from the background and torso, using an automatic parsing method, namely MaskGAN [26]. Similarly to AD-NeRF [19], we assume that the last sample on each ray lies on the background and takes the corresponding RGB color. The neural radiance field representation of the talking head is learned by an implicit function F_Θ, which corresponds to an MLP. Given the talking head at a specific video frame, shown from a particular viewpoint and with a particular expression, we first march camera rays through the scene and sample 3D points on these

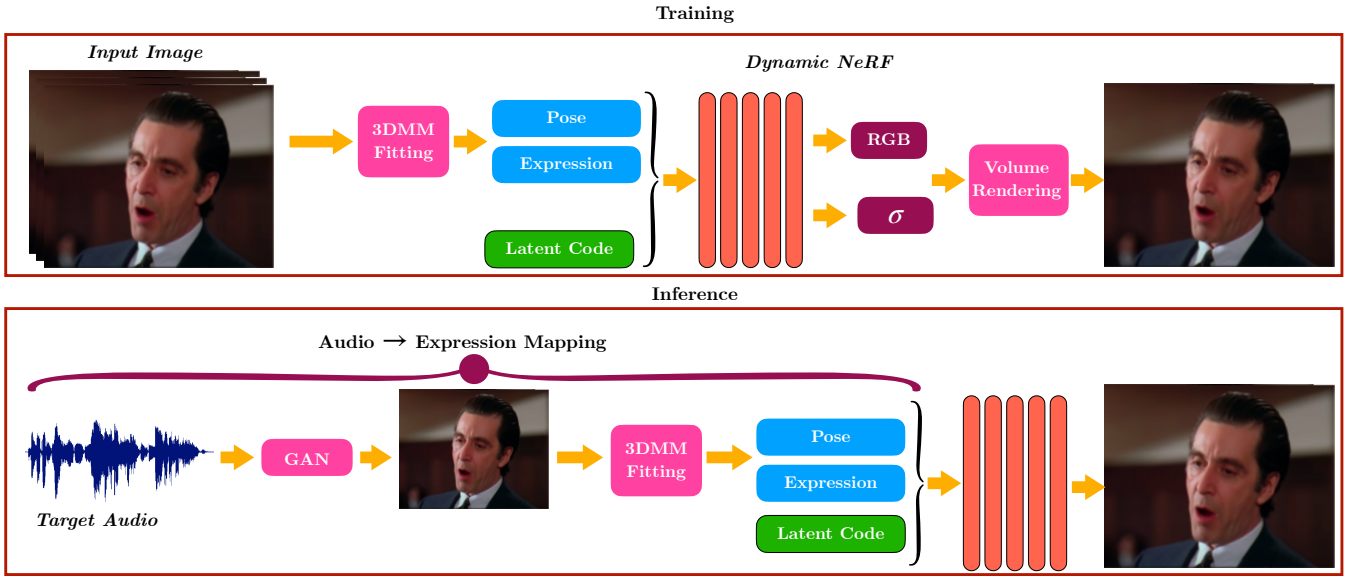

Fig. 2. Overview of our proposed LipNeRF. Given a talking head video, LipNeRF trains a dynamic NeRF conditioned on expression parameters and learned latent codes. During inference, a simple audio-to-expression mapping leads to a high quality lip-synced video for any input audio.

rays. For a 3D point location $\mathbf{x}$, its viewing direction $\mathbf{d}$, the estimated expression parameters $\mathbf{e}$, and a learned latent vector $\mathbf{v}$, F_Θ predicts the RGB color $\mathbf{c}$ and density σ of the point:

$$F_\Theta : (\mathbf{e}, \mathbf{v}, \mathbf{x}, \mathbf{d}) \longrightarrow (\mathbf{c}, \sigma) \tag{1}$$

Given the predicted color $\mathbf{c}$ and density σ for every point on each ray, we can produce the final video frame applying volumetric rendering. For each camera ray $\mathbf{r}(t) = \mathbf{o}+t\mathbf{d}$ with camera center $\mathbf{o}$ and viewing direction $\mathbf{d}$, the color C of the corresponding pixel can be computed by accumulating the predicted colors and densities of the sampled points along the ray:

$$C(\mathbf{r};\Theta) = \int_{t_n}^{t_f} \sigma_\Theta(\mathbf{r}(t))\mathbf{c}_\Theta(\mathbf{r}(t),\mathbf{d})T(t)dt \tag{2}$$

where t_n and t_f are the near and far bounds correspondingly, and $T(t) = \exp\left(-\int_{t_n}^{t}\sigma_\Theta(\mathbf{r}(s))ds\right)$ is the accumulated transmittance along the ray from t_n to t. We denote $\mathbf{c}_\Theta$ and σ_Θ the outputs of F_Θ, omitting the input conditions for short.

Similarly to NeRF [29], we simultaneously optimize a coarse and a fine model with hierarchical volume rendering. Each model is trained to minimize the following objective function:

$$\mathcal{L} = \mathcal{L}_{photo} + \lambda\mathcal{L}_{latent} \tag{3}$$

where $\mathcal{L}_{photo} = \sum_{\mathbf{r}} \left\| \hat{C}(\mathbf{r};\Theta) - C(\mathbf{r};\Theta) \right\|_2^2$ is the photo-consistency loss that measures the pixel-level difference between the ground truth color $C(\mathbf{r};\Theta)$ and the predicted color $\hat{C}(\mathbf{r};\Theta)$ for all the rays $\mathbf{r}$, $\mathcal{L}_{latent} = \|\mathbf{v}\|_2$ regularizes the per-frame latent vectors and λ is set to 0.01.

C. Input Conditions

3D Morphable Model. Given an input video, we first fit a 3DMM on each frame and extract the corresponding head pose and expression parameters. A 3DMM [6] is a parametric model that can represent a face as a linear combination of principle axes for shape, texture, and expression, learned by principal components analysis (PCA). We directly use the learned principal axes from [18], which uses the Basel Face Model (BFM) [31] for shape and texture and the FaceWarehouse [9] for expression. To fit the 3DMM on every video frame, we follow the optimization-based method proposed by [18], that minimizes an objective function with photo-consistency and landmark terms. The extracted head pose corresponds to a 4x4 camera matrix, represented by a rotation matrix $R \in \mathbb{R}^{3x3}$ and a translation matrix $t \in \mathbb{R}^{3x1}$. This is used to transform the sampling points to the canonical space and shoot the rays. The extracted expression coefficients $\mathbf{e} \in \mathbb{R}^{79x1}$ are used as input to the dynamic NeRF.

Learned Latent Codes. In addition to the expression, the dynamic NeRF is also conditioned on per-frame latent codes $\mathbf{v}$. These codes are randomly initialized embeddings that are learned during training. We empirically chose a vector dimension of 32 for each frame, i.e. $\mathbf{v} \in \mathbb{R}^{32x1}$. These vectors can capture the appearance of the talking head at each video frame and memorize characteristics that are independent of the input audio, in order to reconstruct them in the generated video and only modify the lip movements.

D. Lip Synced Video Synthesis

During inference, only the target audio is available, not the lip synced talking head video. Thus, we cannot extract the ground truth expression parameters. Instead, we leverage the lip sync accuracy of a pre-trained Wav2Lip [32] model. Given the input video and target speech segment, we first apply Wav2Lip and fit the 3DMM on its results. Even though Wav2Lip often produces artifacts or blurry results, the 3DMM fitting successfully captures the expression and the lip position at each video frame. The estimated expressions

along with the learned latent codes are passed to the dynamic NeRF, in order to produce the final lip synced video.

We use the per-frame background and torso as given in the source video. In some cases, the synthesized head may be misaligned from the neck, depending on the articulated phoneme in the source and target audio. For example, if the speaker pronounces the phoneme /a/ in the source video (open mouth) and the target phoneme is /k/ (closed mouth), the synthesized result will have a visible gap between the chin and the neck, since those pixels were never visible in the source frame (see Fig. 3). To fill this gap, we inpaint the missing neck region using a generative model [46], [47] trained on CelebA-HQ faces.

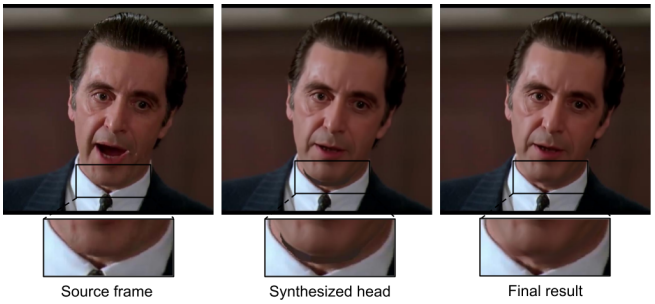

Fig. 3. Head-torso separation. The synthesized head might be separated from the neck, depending on the source and target phonemes. The final result is produced with inpainting.

E. Implementation Details

LipNeRF produces the final talking head video in a frame-by-frame manner. This can result in temporally noisy results. To address this issue, we smooth the expression parameters along the temporal axis. We empirically chose the 1-D Savitzky-Golay FIR filter [35] with window length 5 frames and polynomial order 2. The MLP architecture follows AD-NeRF [19], consisting of 8 linear layers with hidden size of 128 and ReLU activations. Positional encodings are applied to both the 3D locations $\mathbf{x}$ and the viewing directions $\mathbf{d}$ of 10 and 4 frequencies correspondingly. At each iteration, we randomly sample 2048 rays for a video frame. The model is trained for 600k iterations (around 2 days on a single GPU), using Adam optimizer with initial learning rate 5×10^{-4} that decays exponentially to 5×10^{-5}.

IV. EXPERIMENTS

A. Data

Our goal is to synthesize photorealistic lip-synced videos of HD quality, especially in the case of movie dubbing. To the best of our knowledge, there is no publicly available dataset that includes talking head videos from movies with corresponding audio in different languages. Thus, we created a new dataset. We collected 10 videos from popular movies of around 30 seconds to 2 minutes long each, in HD resolution (720p). Along with the original audio in English, we collected and aligned the corresponding dubbed audio for each video in 2 or 3 different languages, including French, Spanish, German and Italian. The videos are sampled at 25 fps and the audio tracks at 16 kHz, with a single channel.

We picked movie scenes where a single actor is speaking and shown by the camera at every moment. We preferred long speeches, e.g. the Al Pacino speech in Scent of a Woman, but we also included scenes where the camera alternates between the target actor and another person or scene. In that case, only the clips of the target actor are used for training. The main data constraints were that the scene and illumination remain almost static throughout each clip. Compared to related works [19], [45], [27], our data are more challenging due to the cinematic lighting (non-uniform lighting of the face), large head pose variations, emphatic head movements, and exaggerated facial expressions.

B. Qualitative Evaluation

We evaluate our method in two cases: (a) original driving audio (reconstruction), and (b) dubbed driving audio (final goal). We compare with several state-of-the-art approaches, namely Speech2Vid [21], MakeItTalk [52], PC-AVS [51], Wav2Lip [32] and AD-NeRF [19]. Please note that MakeItTalk [52] and PC-AVS [51] animate a single face image, so we are mostly interested in their output lip sync accuracy and overall quality. For PC-AVS [51], we use the original video to drive the head pose of the talking head.

Original Audio. Figure 4 illustrates qualitative results for 2 videos from our dataset. Here we use the original English audio as the target speech segment, so we expect accurate reconstruction of the ground truth video (first row). We can clearly see that Speech2Vid suffers from blurry results, MakeItTalk does not produce accurate lip movements and PC-AVS distorts the speaker identity. On the other hand, Wav2Lip produces accurate lip movements but often generates artifacts on the mouth region (see columns 3 and 6). AD-NeRF overfits to the training audio and produces a very accurate reconstruction. Our approach uses the expression parameters extracted from Wav2Lip results, and not from the training video during inference. Yet, it produces a very accurate lip synchronization and good visual quality.

Dubbed Audio. Figure 5 demonstrates qualitative results using the dubbed audio in a different language, which is our final goal. In this case, we do not have ground truth video available. On top of each column we show the target phoneme (or pause). Since Speech2Vid, MakeItTalk and PC-AVS did not perform well even in the simple case of reconstruction, we proceed our comparison with Wav2Lip and AD-NeRF. Wav2Lip, as a GAN trained on thousands of videos, it can generate well-synchronized lip movements to any target audio. However, it frequently produces artifacts on the mouth or teeth, especially in exaggerated expressions and large HD faces (e.g. notice the teeth in top column 2 and bottom column 6, and the black artifacts in the mouth in bottom columns 2, 3, 4). On the other hand, AD-NeRF lacks in terms of lip synchronization. It fails to close the mouth in top column 1 and produce the /δ/ mouth position in top column 4. It also overfits to the training audio, and as a result produces blurry results in the case of dubbed audio,

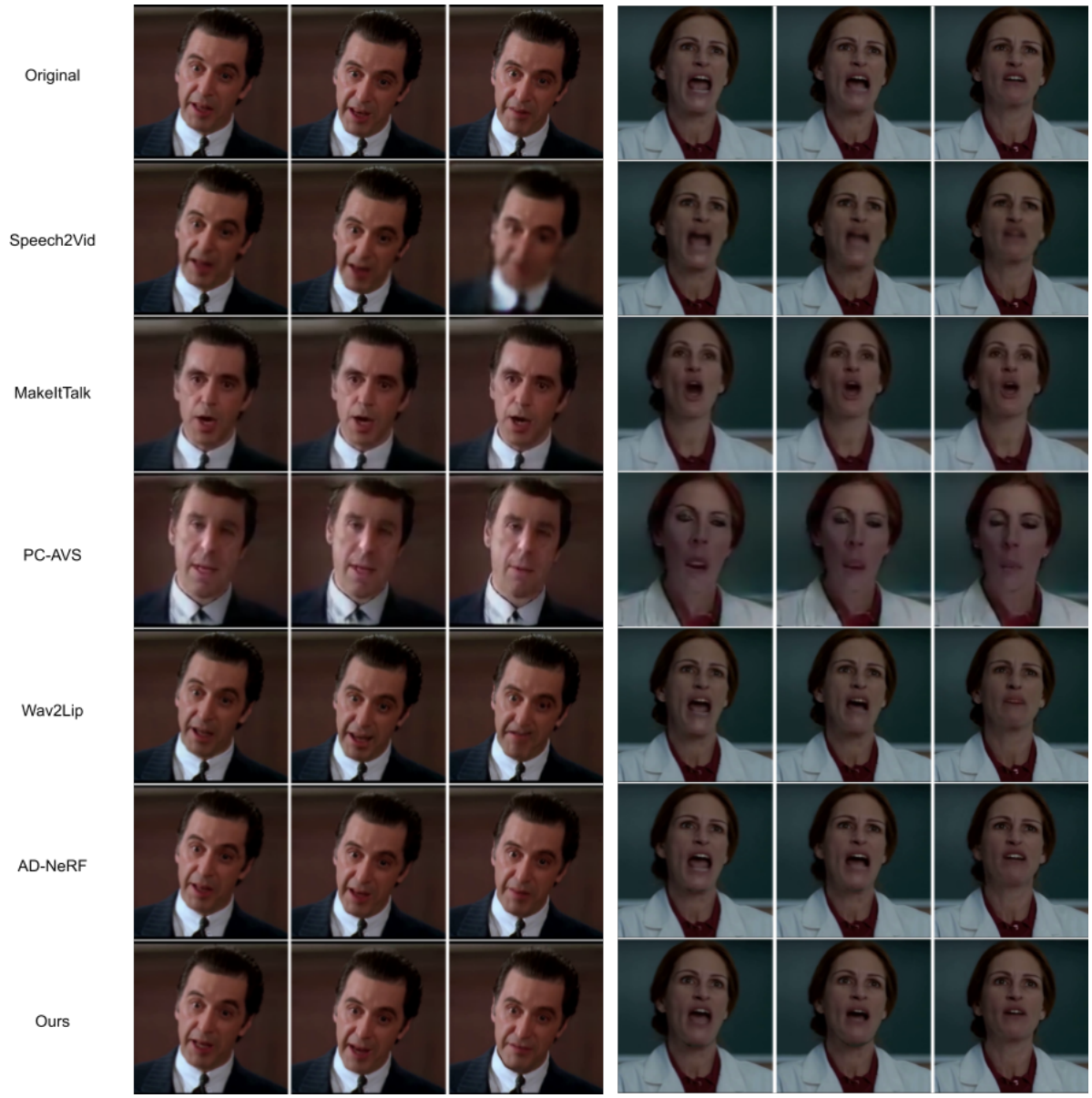

Fig. 4. Qualitative results using the original audio of the source video (reconstruction). First row shows the original video frames (ground truth). Following rows show the corresponding results of Speech2Vid [21], MakeItTalk [52], PC-AVS [51], Wav2Lip [32], AD-NeRF [19], and LipNeRF (Ours).

especially in head movements (see bottom columns 3, 4, 5). In contrast, our method follows the lip sync accuracy of Wav2Lip and generates HD quality. Its superiority over AD-NeRF proves that the expression space is a better representation for lip-syncing NeRFs than the audio feature space. In addition, it avoids any head-torso separation that frequently appears in AD-NeRF (see top columns 5, 6). We strongly encourage readers to watch the supplementary videos in HD.

C. Quantitative Evaluation

Metrics. We quantitatively evaluate the lip synchronization and visual quality of the generated talking head videos. Regarding lip synchronization, we use LSE-D (Lip Sync Error - Distance) and LSE-C (Lip Sync Error - Confidence) metrics [32], based on a pre-trained SyncNet [13]. LSE-D is computed as the average Euclidean distance between corresponding audio and visual embeddings. A lower LSE-D denotes a higher audio-visual match. LSE-C gives a confidence score, estimated as the difference between the minimum and the median Euclidean distance using a sliding window approach [13]. Higher the confidence, the better the audio-video synchronization. To assess the visual quality, we use standard reconstruction metrics, namely peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [44], and learned perceptual image patch similarity (LPIPS) [49].

Results. Table I shows the quantitative results in the dubbed audio case (our final goal) for all the videos in our dataset. LipNeRF demonstrates a significant improvement in lip-sync quality over AD-NeRF and in photorealism over Wav2Lip. Note that Wav2Lip achieves better LSE-D and LSE-C metrics, but it is optimized for these metrics during training. Regarding visual quality, Speech2Vid, MakeItTalk,

Fig. 5. Qualitative results using dubbed audio in Spanish, Italian, and French (see corresponding transcript on top - no ground truth image available). Results shown for Wav2Lip [32], AD-NeRF [19], and LipNeRF (Ours).

and PC-AVS perform very poorly, as also noticed in the qualitative evaluation. Wav2Lip often generates blurry faces or artifacts around the mouth. AD-NeRF overfits to the training audio and lacks in the dubbed audio case. Our method produces results significantly more photorealistic than Wav2Lip [32], without compromising on the quality of lip synchronization as AD-NeRF, demonstrating the superiority of the expression space for lip-syncing.

D. Ablation Study

Clip Length. We also evaluated our method for different clip lengths. Related works [19], [45], [27] train NeRFs on videos of at least 3 minutes duration. We noticed that the expression parameters lead to a better overall quality, especially for shorter videos. Figure 6 shows the LSE-D

TABLE I

QUANTITATIVE RESULTS USING THE DUBBED AUDIO IN ALL AVAILABLE LANGUAGES. THE LAST ROW CORRESPONDS TO THE SOURCE VIDEOS, WHERE THE LIP MOVEMENTS ARE NOT IN SYNC WITH THE DUBBED AUDIO. WE CODE EACH ROW AS BEST, 2ND BEST AND 3RD BEST.

.

Method	LSE-D ↓	LSE-C ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Speech2Vid	10.97	1.71	27.67	0.84	0.26
MakeItTalk	10.23	2.21	14.15	0.25	0.39
PC-AVS	8.78	3.65	14.68	0.30	0.48
Wav2Lip	8.06	4.77	28.90	0.89	0.17
AD-NeRF	11.40	1.28	27.38	0.84	0.20
Ours	9.92	2.71	30.10	0.89	0.16
Dubbed	12.02	0.84	inf	1.0	0.0

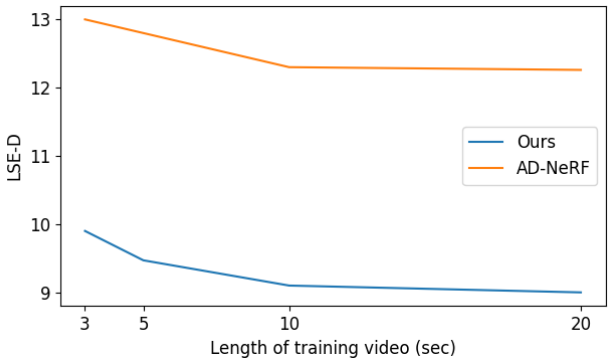

Fig. 6. LSE-D vs training clip length. Results on lip-synced videos with dubbed audio.

Fig. 7. Training clip length 3 sec vs 20 sec. Results using dubbed audio in Spanish.

metric w.r.t. different clip lengths, from only 3 seconds to 20 seconds, in the case of dubbed driving audio. We see that AD-NeRF cannot handle so short clips, whereas our method works well even for as short as 3 seconds. However, we noticed that in the case of emphatic head movements, 3 seconds might not be enough (see artifacts in Fig. 7).

Latent Codes. Here we ablate the contribution of the learned latent codes. If we omit them, we noticed that our method produces similar lip motion and quality but does not memorize some per-frame characteristics. For example, in Figure 8, LipNeRF generates closed eye lids for a dubbed target audio, following the original frame. Without latent codes, the eyes are generated open.

V. DISCUSSION

Limitations. Similarly to other state-of-the-art methods, our method has difficulty synthesizing high-frequency textures, such as facial hair, and geometric details, such as wrinkles (see Supplementary). In those cases, results can be blurry in HD. In addition, dynamic NeRFs, like ours, can handle motion well, compared to standard NeRFs [29] that are designed to represent static scenes. However, results may deteriorate in case of fast and large head movements (see Supplementary). Very short training videos can also lead to blurriness, when lip synced to unseen audio (see Fig. 7).

Future Work. In the future, we plan to address the most challenging cases of pose, motion, and face attribute variations that appear in cinematic videos. We will learn a generic audio-visual representation from multiple identities, in order to densely sample phonemes and visemes from a wide variety of speaker-specific characteristics. Regarding high-frequency textures, a possible solution would be to learn a multiscale representation, such as Mip-NeRF [5].

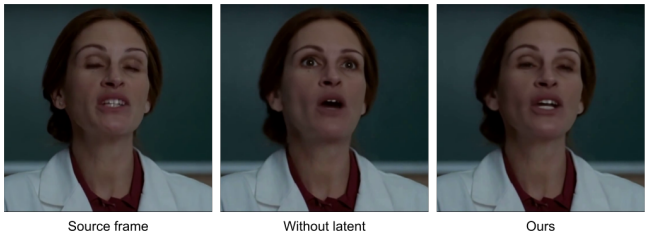

Fig. 8. Ablation study of the learned latent vectors.

Ethical Considerations. We would like to note the potential misuse of video synthesis methods. With the rise of "deep fakes", it becomes easier to generate photorealistic fake videos of any speaker. These can be used for malicious purposes, e.g. to spread misinformation. To this end, it is important to develop accurate methods for fake content detection and forensics [48], [43]. In addition, appropriate procedures must be followed to ensure fair and safe use of videos if used for training or inference.

VI. CONCLUSION

In conclusion, we propose LipNeRF, a method for audio-driven talking head video synthesis that outperforms current state-of-the-art methods. LipNeRF learns a dynamic NeRF conditioned on the expression space of a 3DMM, instead of the audio feature space. During inference, we propose a simple audio-to-expression mapping, leveraging the lip sync accuracy of a pre-trained GAN-based model. Compared to audio-conditioned NeRFs, LipNeRF produces better lip synchronization and visual quality, especially in the case of unseen audio. To evaluate our method, we collected an HD dataset of talking heads from movie scenes, with corresponding dubbed audio in different languages. To the best of our knowledge, this is the first dataset designed for the purpose of movie dubbing. LipNeRF is able to accurately model the 3D face geometry, handling challenging pose, illumination and facial expression variations. We quantitatively and qualitatively demonstrate the superiority of our method over the current state-of-the-art.

Acknowledgements. This work was supported by Amazon Prime Video, Partner University Fund 4DVision Project, and SUNY2020 Infrastructure Transportation Security Center.

REFERENCES

[1] S. Athar, A. Pumarola, F. Moreno-Noguer, and D. Samaras. Facedet3d: Facial expressions with 3d geometric detail prediction. *arXiv preprint arXiv:2012.07999*, 2020.
[2] S. Athar, Z. Shu, and D. Samaras. Self-supervised deformation modeling for facial expression editing. 2020.
[3] S. Athar, Z. Shu, and D. Samaras. Flame-in-nerf: Neural control of radiance fields for free view face animation. *arXiv preprint arXiv:2108.04913*, 2021.
[4] S. Athar, Z. Xu, K. Sunkavalli, E. Shechtman, and Z. Shu. Rignerf: Fully controllable neural 3d portraits. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.
[5] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields, 2021.

[6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[7] M. Brand. Voice puppetry. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 21–28, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[8] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997.

[9] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.

[10] L. Chen, R. K. Maddox, Z. Duan, and C. Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019.

[11] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

[12] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020.

[13] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016.

[14] M. Doukas, M. R. Koujan, V. Sharmanska, A. Roussos, and S. Zafeiriou. Head2head++: Deep facial attributes re-targeting. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3:31–43, 2021.

[15] B. Fan, L. Wang, F. K. Soong, and L. Xie. Photo-real talking head with deep bidirectional lstm. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888, 2015.

[16] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. K. Kakumanu, and O. N. Garcia. Audio/visual mapping with cross-modal hidden markov models. *IEEE Transactions on Multimedia*, 7(2):243–252, 2005.

[17] G. Gafni, J. Thies, M. Zollhöfer, and M. Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction, 2020.

[18] Y. Guo, J. Cai, B. Jiang, J. Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1294–1307, 2018.

[19] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. 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.

[20] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. Deep speech: Scaling up end-to-end speech recognition, 2014.

[21] A. Jamaludin, J. S. Chung, and A. Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127, 12 2019.

[22] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Trans. Graph*, 2018.

[23] T. Kim, Y. Yue, S. Taylor, and I. Matthews. A decision tree framework for spatiotemporal sequence prediction. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 577–586, 2015.

[24] M. Koujan, M. Doukas, A. Roussos, and S. Zafeiriou. Head2head: Video-based neural head synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 319–326, Los Alamitos, CA, USA, may 2020. IEEE Computer Society.

[25] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2755–2764, June 2021.

[26] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

[27] X. Liu, Y. Xu, Q. Wu, H. Zhou, W. Wu, and B. Zhou. Semantic-aware implicit neural audio-driven video portrait generation. *arXiv*, abs/2201.07786, 2022.

[28] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021.

[29] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[30] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021.

[31] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[32] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. 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*, page 484–492, 2020.

[33] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: One-shot anatomically consistent facial animation. *International Journal of Computer Vision*, 128(3):698–713, 2020.

[34] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Hmm-based text-to-audio-visual speech synthesis. In *Sixth International Conference on Spoken Language Processing*, 2000.

[35] A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[36] T. Shimba, R. Sakurai, H. Yamazoe, and J.-H. Lee. Talking heads synthesis from audio with deep neural networks. *2015 IEEE/SICE International Symposium on System Integration (SII)*, pages 100–105, 2015.

[37] Z. Shu, M. Sahasrabudhe, R. Alp Guler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018.

[38] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017.

[39] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[40] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, pages 716–731. Springer, 2020.

[41] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering. *ACM Trans. Graph*, 2019.

[42] K. Vougioukas, S. Petridis, and M. Pantic. End-to-end speech-driven realistic facial animation with temporal gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[43] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[45] S. Yao, R. Zhong, Y. Yan, G. Zhai, and X. Yang. DFA-NeRF: Personalized Talking Head Generation via Disentangled Face Attributes Neural Rendering. *arXiv*, abs/2201.00791, 2022.

[46] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[47] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892*, 2018.

[48] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *ICCV*, 2021.

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[50] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. 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.

[51] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[52] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li. Makeittalk: Speaker-aware talking-head animation. *ACM Transactions on Graphics*, 39(6), 2020.