What comprises a good talking-head video generation?

Lele Chen Guofeng Cui Ziyi Kou Haitian Zheng Chenliang Xu
University of Rochester
lchen63@ur.,gcui2@ur.,zkou2@ur.,hzheng15@ur.,chenliang.xu@}rochester.edu

1. Introduction

Over the years, performance evaluation has become essential in computer vision, enabling tangible progress in many sub-fields. While talking-head video generation has become an emerging research topic, existing evaluations on this topic present many limitations. For example, most approaches use human subjects (e.g., via Amazon MTurk) to evaluate their research claims directly. This subjective evaluation is cumbersome, unreproducible, and may impend the evolution of new research. In this work, we present a carefully-designed benchmark for evaluating talking-head video generation. By conducting a thoughtful analysis across several state-of-the-art talking-head generation approaches, we aim to uncover the merits and drawbacks of current methods and point out promising directions for future work. A full version of this survey can be found at here.

A sizable volume of follow-up papers have been published since the introduction of identity-independent talking-head generation task [9]. While there has been substantial progress in terms of synthesized video quality, relatively less effort has been spent in evaluating talking-head methods, and grounded ways to quantitatively assess these videos are still missing. While some existing metrics are shown to be effective image-level visual quality evaluation, there are some other issues, such as the variety of probability criteria and the lack of perceptually meaningful video-level measures, have made evaluating the talking-head video generative models notoriously tricky. In this paper, we mainly discuss and assess talking-head video generative approaches by either designing or choosing evaluation metrics concerning the three desiderata:

Identity Preserving. We compare two existing identity-preserving evaluation metrics by visualizing the decision boundaries of inter-class discrepancy ability, and select cosine similarity between embedding vectors of ArcFace [6] to measure identity mismatch.

Semantic-level Lip Synchronization. While some methods can generate realistically looking videos, the generated lip movements usually present less expressive and discriminative semantic cues, which can not convey the audio information. To address this semantic lip-synchronization ability, we critically discuss existing lip-sync evaluation methods and introduce a new lip-sync metric—lipreading Similarity Distance (LRSD), which evaluates the lip movement synchronization in semantic perspective. The experimental results demonstrate that our LRSD score agrees with human perceptual judgments and human rankings of videos.

Natural-spontaneous Motion. Video generative models have well-known limitations, including a tendency towards limited diversity in generated video samples. In order to investigate intra-video diversity, we evaluate the spontaneous motions emitted in synthesized videos, including emotional expression, blinks, and head movements. Meanwhile, we design a new evaluation metric—Emotion Similarity Distance (ESD) to evaluate the facial emotional expression distance between the synthesized video and the ground truth. To quantitatively evaluate the subconscious blinks in a talking-head video, we introduce a learning-based metric—Blink Similarity Distance (BSD)to evaluate the quality of the blink motion in the eye region of a synthesized video.

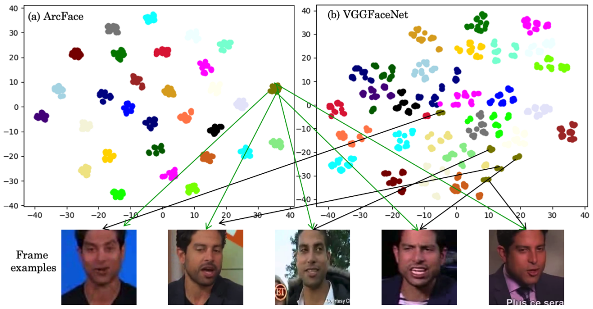

Figure 1. The t-SNE plot of identity features of random frames from VoxCeleb2 testing set. The features are extracted by VGFaceNet and ArcFace, respectively. Frames corresponding to the same subject have the same color.

2. Identity Preserving

People are sensitive to any perceptual identity changes in a synthesized video, is hard to avoid in the deep generative model. The reason is that the spatial identity information may not be preserved perfectly after deep convolution layers (e.g., encoding and fusion network).

To evaluate the identity-preserving performance, Jamaludin et al. [9] use the embedding distance of the generated video frames and the ground truth using a pre-trained VGGFaceNet [13] to measure the identity distance since it is trained with a triplet loss. ArcFace [6] has been adopted in Zakharov et al. [20].

To compare those two different embedding methods, we use t-SNE [12] to visualize the extracted feature vectors of video frames sampled from VoxCeleb2 (see Fig. 1). From the t-SNE plot, we find that ArcFace (Fig. 1a) is more robust to noise (e.g., hairstyle, lighting, and video quality) comparing with VGGFaceNet (Fig. 1b). We attribute this to the Additive Angular Margin Loss (ArcFace) [6] since it simultaneously enhances the intra-class compactness and inter-class discrepancy. Based on the observation that ArcFace has better inter-class discrepancy ability, we use ArcSim—the cosine distance between the two image features extracted by ArcFace to measure the identity similarity between two images.

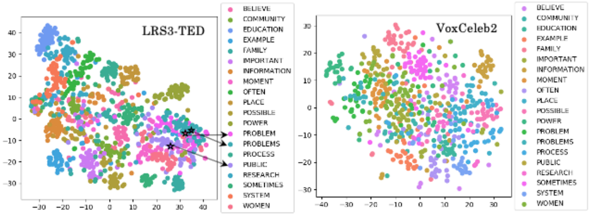

Figure 2. The t-SNE plots of semantic-level visual features of random videos from the LRS3-TED testing set and VoxCeleb2 testing set. Videos corresponding to the same word have the same color.

3. Semantic-level Lip Synchronization

Another challenge of talking-head generation is to maintain the synchronization between visual dynamics (e.g., facial movement, lip movement) and the driven modality (e.g., audio signal, and landmark) since people are sensitive to the slight misalignment between facial movements and speech audio. When humans look at a talking-head video, we would unintentionally use the semantic information to judge if the audio is synced with the visual. For example, it is easier for us to tell if the audio is synced with the visual when we know the language. Thus, in this paper, we propose a lip-synchronization evaluation metric—Lip-Reading Similarity Distance (LRSD), like human perceptual judgments by incorporating the semantic-level lipreading. Given a synthesized video clip $\hat{x}$ and paired ground truth video clip x, the LRSD is obtained by: LRSD$(x, \hat{x}) = ||\phi(x), \phi(\hat{x})||_2^2$, where the ϕ is a spatial-temporal lipreading network. Although there are many lipreading networks proposed in recent years, most of them do not perform well on videos outside the dataset, not to mention assessing the similarity between synthesized videos and real videos. Thus, we propose an easy but effective multi-view lipreading network, which is trained on LRS3-TED dataset and works for any video outside the dataset.

To demonstrate the visual feature extraction ability of our lipreading network, we show the lipreading results on the testing set of LRS3-TED and VoxCeleb2 datasets in Fig. 2. In order to show the inter-class discrepancy ability of the lipreading feature, we randomly select 20 words from our vocabulary, each with 30 video clips in each testing set, and visualize their visual features. We can find that the features of words with similar visemes are closer than other features. We also show the lipreading accuracy in Tab. 3, from where we can find that our lipreading network achieves 42.46% top-1 classification accuracy on the VoxCeleb2 testing set (note that our classifier is trained on LRS3-TED dataset). From the t-SNE plot and classification accuracy, we can find that our lipreading network can extract the semantic-level spatial-temporal features from the input video sequence, and there are clear margins between the extracted features when they do not belong to a same word.

4. Natural-spontaneous Motions

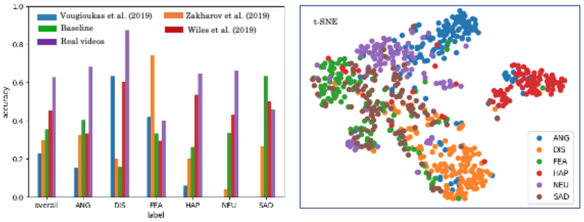

Figure 3. The left side shows the video emotion classifier's performance on the CREMA-D testing set. The X-axis and Y-axis are emotion labels and classification accuracy, respectively. The right side is the t-SNE plot of the emotion encoding of random video samples from the CREMA-D testing set. Videos corresponding to the same emotion label have the same color.

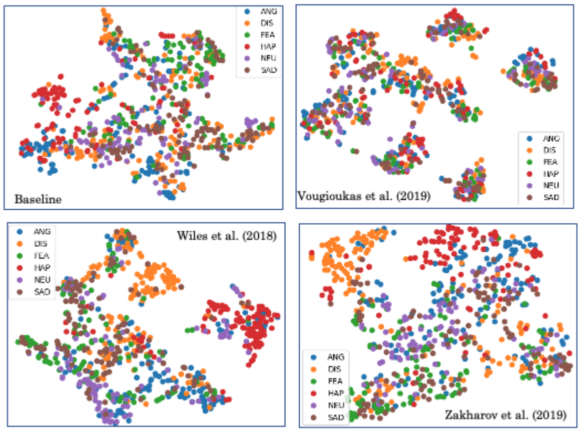

Figure 4. The t-SNE plot of the ESD features of synthesized videos produced by different methods.

People naturally emit spontaneous motions such as head movements and emotional expressions when they speak, which contain nonverbal information that helps the audience comprehend the speech content [3, 8]. Although speech contains necessary information for generating lip movements, it can hardly be used to produce natural-spontaneous motions. Some works [7, 9, 14, 5] ignore the modeling of spontaneous expressions, resulting in faces that are mostly static except for the mouth region. Karras et al. [10] propose a network to synthesize 3D vertex by inferring the information from the audio signal and emotional state. Vougioukas et al. [16] propose a noise generator capable of producing noise that is temporally coherent through a single-layer GRU. This latent representation introduces randomness in the face synthesis process and helps with the generation of blinks and brow movements. Some works [19, 15, 11, 1, 21] take image frames that contain the target motion as dense mapping to guide the video generation, producing video frames with convoluted head motion and facial expressions.

Investigating the generation of spontaneous expressions is also important since it is one of the main factors that affect our perception of how natural a video looks. To evaluate the quality of synthesized spontaneous motion (emotional expression), we introduce a new emotion similarity distance (ESD). We first train a spatial-temporal convolution network to classify emotions of video clips in the CREMA-D training set. The left side of Fig. 3 shows the video emotion classification accuracy on the CREMA-D testing set. According to the user studies in Cao et al. [2], the human recognition of intended emotion on the CREMA-D dataset are 58.2% (visual-only) and 63.6% (audio-visual), respectively. Our video-level emotion classifier achieves 62.9% testing accuracy on real videos without audio, which is better than individual human raters (visual-only). Then we training the network with the metric learning step using ArcLoss, which leads

	LRS3-TED						VoxCeleb2					
	Real	Chen [5]	Baseline	Wang [17]	Wiles [18]	Zakharov [20]	Real	Chen [5]	Baseline	Wang [17]	Wiles [18]	Zakharov [20]
Top-1	72.62%	1.99%	**3.85%**	2.23%	1.93%	1.77%	42.46%	2.40%	**3.04%**	1.64%	1.87%	1.99%
Top-5	87.98%	5.01%	**11.05%**	6.19%	5.73%	5.48%	63.98%	**8.12%**	7.13%	4.56%	4.80%	4.74%
Top-10	91.53%	8.19%	**16.13%**	8.37%	8.62%	8.27%	70.82%	**10.76%**	**10.76%**	7.13%	7.49%	7.54%
Top-20	94.42%	12.99%	**22.11%**	12.27%	13.18%	12.53 %	78.30%	15.80%	**15.91%**	11.23%	11.70%	11.05%
LRSD	—	46.35 %	**59.60%**	56.25%	52.95%	51.93%	—	47.93%	**62.56%**	61.59%	55.14%	53.87%
L2	—	1.03	**0.89**	0.93	0.96	0.98	—	1.02	**0.86**	0.87	0.94	0.96

Table 1. Semantic-level video quality of different methods. The $L2$ row shows the $L2$ distance between features of the fake video and the paired real video.

Method	Baseline	Zakharov [20]	Vougioukas [16]	Chen [4]	Jamaludin [9]	Wang [17]	Wiles [18]
BSD	0.965	0.935	0.919	0.907	0.807	0.957	**0.979**

Table 2. The BSD score of different methods on Grid testing set. BSD score measures cosine similarity between blink features extracted from synthesized video and ground truth video.

Method	Baseline	Zakharov [20]	Wiles [18]	Vougioukas [16]
ESD ↑	0.467	0.391	**0.655**	0.2665

Table 3. The ESD score across over different methods on CREMA-D testing set. We bold the leading score.

to a more clear margin between different emotion features. We extract the emotion features from real videos in CREMA-D testing set and show the t-SNE plot in Fig. 3 right side. Since the video features from our model are optimized on a hypersphere with cosine angles, it is naturally to apply cosine similarity as Emotion Similarity Distance (ESD). Therefore, after training with ArcLoss, we utilize the embedding features before the ArcLoss module to represent each input video and calculate their similarity distance as:

$$\mathrm{ESD}(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} = \frac{\sum_{k=1}^{n} v_i^k v_j^k}{\sqrt{\sum_{k=1}^{n} (v_i^k)^2} \sqrt{\sum_{k=1}^{n} (v_j^k)^2}} , \tag{1}$$

where i and j are indexes for two videos respectively. The ESD result is shown in Tab. 3 and we will discuss it in the following part.

Fig. 3 left side shows the emotion classification accuracy on different types of videos. Fig. 4 shows the t-SNE plot of different ESD features on CREMA-D testing set, from where we observe that the group boundaries of ESD feature extracted from baseline method, Wiles et al. [18], and Zakharov et al. [20] are more clear than the ESD feature extracted from synthesized videos produced by Vougioukas et al. [16]. The t-SNE visualization is consistent with the classification results in the first row of Fig. 4, where the emotion classifier achieves lowest accuracy on synthesized videos produced by Vougioukas et al. [16]. Tab. 3 shows the quantitative result of ESD, from where we can find that the emotional feature extracted from Wiles et al. [18] is closest to the feature extracted from ground truth comparing to other methods. This is consistent with the emotion classification accuracy shown in Fig. 3 second row, where the synthesized videos produced by Wiles et al. [18] achieves highest classification accuracy (45.3%). In summary, the results shown in Fig. 3, Fig. 4, and Tab. 3 demonstrate that our ESD is a well-characterized perceptual similarity measure that aims to assess the emotional expression ability of synthesized videos.

We also train the network with ArcLoss and extract blink features for each slices the same as what we do for emotion features. The t-SNE plot over blink features of sampled slices from test set is shown in Fig. 5. Although confusion slices exist for blink model, the blink features represent obvious inter-class discrepancy ability, that is non-blink motion cluster on the left and blink motion cluster on the right. Based on this observation, we introduce Blinking Similarity Distance (BSD) to better evaluate blink generation quality of synthesized videos. Similar to ESD, we calculate the cosine similarity between blink feature of ground truth videos and that of synthesized videos (same equation as Eq. 1). A high score of BSD indicates similarity blink motion between two videos, which means both of them perform either similar blink motion or similar non-blink motion.

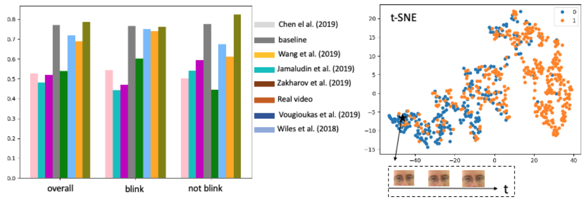

Figure 5. The evaluation of blink motion. The histogram shows the performance of blink model on GRID's testing set. We evaluate the model on both original real videos in GRID and also synthesized videos by seven example methods. We also show the t-SNE plot for blink features extracted by our blink model. In the bottom figure, points with label 1,0 are belonged to blink motion, and non-blink motion, respectively.

5. Conclusion

Talking-head generation is an important and challenging problem in computer vision and has received considerable attention. Thanks to remarkable developments in GAN techniques, the field of talking-head generation has dramatically evolved. we introduced three perceptually meaningful metrics that assess the emotional expression, semantic-level lip synchronization, and blink motion of a synthesized video. The proposed metrics agree with human perceptual judgment, and have low sample and computational complexity. The performance of talking-head generation will continue to improve as various structures are proposed. In the mean time, seeking appropriate measures for this task continues to be an important open problem, not only for fair model comparison but also for understanding, improving, and developing the talking-head animation models.

Acknowledgments. This work was supported in part by NSF 1741472, 1813709, and 1909912. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen. Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, 36(6):196, 2017.

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

[3] J. Cassell, D. McNeill, and K.-E. McCullough. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & cognition*, 7(1):1–34, 1999.

[4] L. Chen, Z. Li, R. K Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018.

[5] 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 Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[7] 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. IEEE, 2015.

[8] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019.

[9] A. Jamaludin, J. S. Chung, and A. Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127(11-12):1767–1779, 2019.

[10] T. Karras, T. Aila, S. Laine, A. Herva, and J. 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.

[11] 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 Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[12] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In M. W. J. Xianghua Xie and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[14] Y. Song, J. Zhu, X. Wang, and H. Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018.

[15] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[16] K. Vougioukas, S. Petridis, and M. Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, pages 1–16, 2019.

[17] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019.

[18] O. Wiles, A. Sophia Koepke, and A. Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–686, 2018.

[19] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu. Audio-driven talking face video generation with natural head pose. *arXiv preprint arXiv:2002.10137*, 2020.

[20] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9459–9468, 2019.

[21] Y. Zhang, S. Zhang, Y. He, C. Li, C. C. Loy, and Z. Liu. One-shot face reenactment. *arXiv preprint arXiv:1908.03251*, 2019.