Deeply Emotional Talking Head: A Generative Adversarial Network Approach to Expressive Speech Synthesis with Emotion Control

Filipe Antonio de Barros Reis
Dept. of Computer Engineering and Industrial Automation
School of Electrical and Computer Engineering
University of Campinas (UNICAMP)
Campinas, Brazil

Paula Dornhofer Paro Costa
Dept. of Computer Engineering and Industrial Automation
School of Electrical and Computer Engineering
University of Campinas (UNICAMP)
Campinas, Brazil

José Mario de Martino
Dept. of Computer Engineering and Industrial Automation
School of Electrical and Computer Engineering
University of Campinas (UNICAMP)
Campinas, Brazil

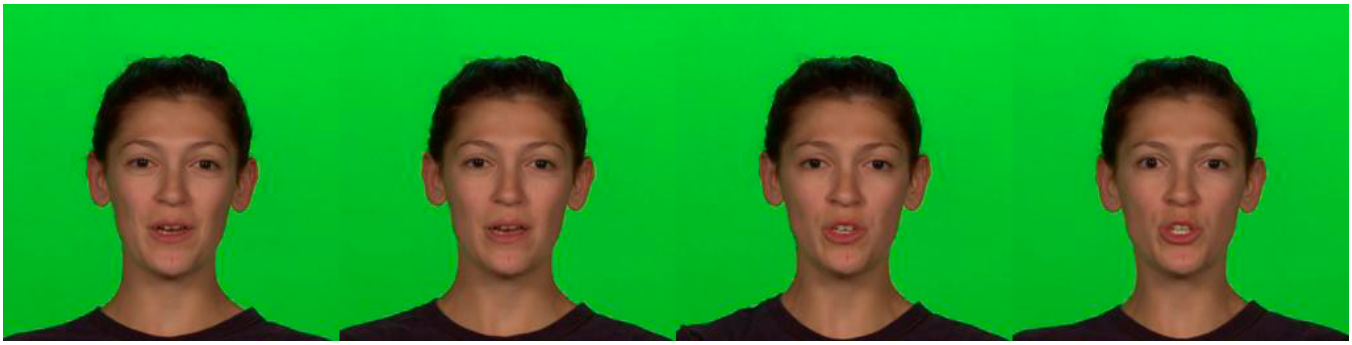

Figure 1: Synthesized images for different emotions generated using the proposed method,which is a novel approach to talking-head synthesis using Generative Adversarial Networks and dedicated structures for emotion control. Given facial keypoints and the desired emotion, our system is able to synthesize expressive talking-head video.

ACM Reference Format:
Filipe Antonio de Barros Reis, Paula Dornhofer Paro Costa, and José Mario de Martino. 2020. Deeply Emotional Talking Head: A Generative Adversarial Network Approach to Expressive Speech Synthesis with Emotion Control. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '20 Posters), August 17, 2020.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3388770.3407417

1 INTRODUCTION

The recent development in natural language processing allowed the widespread use of voice-based virtual assistants on various tasks, ranging from personal assistants capable of helping on simple tasks to customer-facing assistants capable of understanding and solving personal issues with a given service. Although these assistants are capable of completing the task assigned, their interaction with humans still lacks many resources that humans adopt to communicate effectively. Speech is naturally multimodal and contains a non-verbal part since the structures used to generate sounds produce visible results during the speech, which can be interpreted and add information. In particular, the visual information carried by the face during speech is a crucial aspect of communicating emotion and improving the intelligibility of a message [Mattheyses and Verhelst 2015].

Some of these non-verbal resources are obtained through facial expressions and mouth articulation during the speech. Adding embodiment to the assistants is one viable approach to improve communication, as this improves the intelligibility of the communication [Mattheyses and Verhelst 2015]. Additionally, adding emotion to the virtual assistants may generate empathy and make users more engaged with the assistant [Fraser et al. 2018].

To allow the control of emotions portrayed by the actor's facial expression, [Ma and Deng 2019] proposed an approach that allows changing the facial expression of a pre-recorded video without additional driving sources. This approach is relevant because it allows the change of facial expressions without recording more samples or post-processing. One downside of this approach is the need to have all the desired sentences pre-recorded.

To avoid having all the sentences pre-recorded, some authors have proposed face puppeteering systems [Wang et al. 2018; Zakharov et al. 2019]. These systems are capable of generating videos of a reference actor speaking new sentences by using a set of facial keypoints as input. The training of such systems is performed by using a set of pre-recorded videos from a given actor. After the training stage is completed, the only input needed for this system is a set of keypoints. These systems are suitable for use with

SIGGRAPH '20 Posters, August 17, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7973-1/20/08.
https://doi.org/10.1145/3388770.3407417

Table 1: Results for the videos synthesized using our proposed network and the original *vid2vid* network. These results show the potential that our approach has to improve the videorealism of the synthesized videos.

Approach	User Group Preference (%)	FID
Vid2vid	45.3	38
Our Approach	54.7	32

Table 2: Results for the emotion perception test of the study group evaluation. The videos presented were synthesized using keypoints obtained from a neutral emotion, *resentment*, which was not part of the training set.

Emotion	Correct Emotion (%)	Correct Valence (%)	Incorrect Valence (%)
Happy-for	81.7	17.1	1.2
Admiration	35.4	63.4	1.2
Fear	74.4	18.3	7.3
Anger	97.6	0	2.4

virtual assistants as the input keypoints may be obtained either from a reference actor speaking the desired sentences or from a model capable of translating audio or text into facial keypoints. One downside of such systems is that the emotion related to the facial expression synthesized is given only by the combination of the source keypoints, not allowing any control over the desired output.

We propose a face puppeteering system that also implements control over the emotion expressed on the resultant video. We achieve this by expanding the *vid2vid* system proposed by [Wang et al. 2018], and adding components that allow the control over the emotion associated with the facial expression on the output video.

Our main objective in this work is to generate expressive visual speech capable of conveying a target emotion through facial expression. We conducted a subjective perceptual evaluation to determine the performance of our method. To produce the results for this evaluation, we have used a training set composed of a total of 166 seconds of video of an actress playing four different emotions of the *Ortony, Clore and Collins's (OCC)* emotion model, "happy for", "admiration", "fear" and "anger". The evaluation results demonstrate that the videos synthesized using our method results are considered more realistic than those obtained using *vid2vid*, which is used as a reference for this work. Additionally, the users could perceive the correct emotion valence on the majority of the interactions and could also recognize the real target emotions.

2 OUR APPROACH

Our approach to expressive visual speech synthesis expands on the *vid2vid Generative Adversarial Network (GAN)* system, proposed by [Wang et al. 2018]. We propose the addition of a multi-scale PatchGAN discriminator to assess the emotion perceived on the synthesized images and evaluate if it is the targeted emotion. We also propose the inclusion of emotion information on the segmentation maps used as input. The original segmentation maps contained only the drawing of the face delimited by the facial keypoints. We propose the association of the background and line colors with the targeted emotion, to use more information as input without changing the complexity of the network.

To test our proposal, we have used part of the CH-Unicamp dataset [Costa 2015] as our training dataset. This dataset is composed of videos of an actress speaking phrases designed to include the most relevant context-dependent visemes for the Brazilian Portuguese language while performing the emotions of the *Ortony, Clore and Collins's* emotion model.

To evaluate our results, we have performed a user group study with 42 test subjects. We asked the participants to indicate which emotion they perceived in the videos generated with our approach. Additionally, we asked the users to evaluate which video seemed more realistic, one obtained with our approach or another generated using *vid2vid*.

In Table 1 we present the results for the preference test, and for the *Fréchet Inception Distance (FID)* score [Heusel et al. 2017]. This score is an objective metric capable of determining how realistic are the synthesized frame. In both *FID* score and the user preference, our approach outperforms *vid2vid*. This result shows that our network has the potential to generate more realistic results while adding control over the emotion associated with the facial expression presented in the output.

Recognizing different emotions in the *OCC* model is a challenging task for users, even when considering real videos [Costa 2015]. The results for the emotion perception test are presented in Table 2. These results demonstrate that the users are mostly able to perceive the target emotion on the synthesized videos, even though this is a challenging task. Additionally, when the users perceive a different emotion than the target one, it is usually an emotion with the same valence as the target emotion, which may be linked to the difficulty to distinguish between the *OCC* emotions within the same valence.

‘

REFERENCES

Paula Costa. 2015. *Two-Dimensional Expressive Speech Animation.* Ph.D. Dissertation. Universidade Estadual de Campinas. https://doi.org/10.13140/RG.2.1.3131.6968

Jamie Fraser, Ioannis Papaioannou, and Oliver Lemon. 2018. Spoken Conversational AI in Video Games: Emotional Dialogue Management Increases User Engagement. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 179–184. https://doi.org/10.1145/3267851.3267896

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Guenter Klambauer, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. *ArXiv* abs/1706.08500 (2017).

L. Ma and Z. Deng. 2019. Real-Time Facial Expression Transformation for Monocular RGB Video. *Comput. Graph. Forum* 38 (2019), 470–481.

Wesley Mattheyses and Werner Verhelst. 2015. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication* 66 (Feb. 2015), 182–217. https://doi.org/10.1016/j.specom.2014.11.001

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-Video Synthesis. *CoRR* abs/1808.06601 (2018). arXiv:1808.06601 http://arxiv.org/abs/1808.06601

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. *ArXiv* abs/1905.08233 (2019).