Eurographics / ACM IMX Joint Workshop WICED x Cinemotions (2023)
H.-Y. Wu, S. Jolly, A. Bruckert and L. Lévêque (Editors)

Is this the real life? Investigating the credibility of synthesized faces and voices created by amateurs using artificial intelligence tools.

Daniel Gregory[1], Diego Monteiro[2*]

[1] Birmingham City University, B4 7XG, Birmingham, United Kingdom
[2] ESIEA, 53000, Laval, France
* *Corresponding Author.*

Abstract

The widespread availability and accessibility of artificial intelligence (AI) tools have enabled experts to create content that fools many and needs deep scrutiny to be discernible from reality; nevertheless, it is unclear whether presented with the same tools amateurs can also create synthesized faces and voices with similar ease. The possibility of creating this kind of content can be life-changing for smaller movie makers.
Thus it is important to understand how, can amateurs be supported and guided into creating similar media and how believable are their results. This paper aims to propose a framework that can be used by amateurs to create completely artificial content and investigate the credibility of synthesized faces and voices created by amateurs using AI tools. Specifically, we explore whether an entirely AI-generated piece of media, encompassing both visual and audio components, can be convincingly created by non-experts.
To achieve this, we conducted a series of experiments in which participants were asked to evaluate the credibility of synthesized media produced by amateurs. We analyzed the responses and evaluated the extent to which the synthesized media could pass as authentic to the participants.
Our findings suggest that, while AI-generated media created by amateurs may appear visually convincing, the audio component is still lacking in terms of naturalness and authenticity. However, we also found that participants' perceptions of credibility were influenced by their prior knowledge of AI-generated media and their familiarity with the source material.
Our findings also suggest that while AI-generated media has the potential to be highly convincing, current AI tools and techniques are still far from achieving perfect emulation of human behavior and speech, when done by amateurs without artistic interference.

Keywords
Deep Fake, Automatic Generation, Movies, User Study

CCS Concepts
• ***Applied computing*** → ***Arts and humanities;*** • ***Computing methodologies*** → *Artificial intelligence;*

1. Introduction

Artificial intelligence (AI) has made significant progress in recent years, enabling the creation of synthesized media like deepfake technologies that facilitate the manipulation of faces in videos. AI can generate footage of virtually anyone with an online presence, making them appear to say anything. This poses challenges for determining the credibility of media content.

AI technologies can not only replace faces in videos but also generate non-existent faces[†], synthesize speech from text[‡], and even produce text based on prompts[§]. These capabilities render AI a valuable tool for video content creators, as it can expedite the process and reduce costs compared to traditional methods involving hiring actors, crew, and equipment. Theoretically, with these technologies, anyone should be able to create believable scenes. Never-

[†] thispersondoesnotexist.com/
[‡] descript.com/
[§] sassbook.com/ai-writer

theless, malicious use of these technologies could lead to the spread of false information and ensuing chaos.

This paper delves into several themes that collectively explore the process of creating artificially generated media. One such theme is artificial face generation. To create a deepfake using solely artificial elements, an artificial face must first be generated. Keywords related to this theme include StyleGAN, a leading software for generating artificial faces, ALAE, an alternative approach, and other terms such as 'generator' and 'artificial intelligence.'

We aim to elucidate factors influencing the credibility of AI-generated faces and voices and offer valuable information for individuals considering AI for similar projects or those seeking to understand its limitations. We will focus on the entirety of AI-generated video content and evaluate its persuasiveness for a general audience.

The primary objective of this project is to generate multiple deepfakes incorporating synthetic audio and verify how believable this content is. To achieve this, we employed various research methodologies based on their quality and ease of use. These selected methods were combined to create media samples, which were then incorporated into a survey administered to random, anonymous participants to ensure unbiased and impartial results.

The study aimed to achieve the main objective of investigating the development of artificially created media. Three parts of the development were studied, including audio synthesis, deepfake creation, and video generation. Audio synthesis involves synthetic voices, deepfake creation uses deepfake technology to create an artificial face, and video generation creates various elements that do not require a person in the frame. Relevant keywords for these themes include voice cloning, artificial intelligence, DeepFaceLab for deepfake creation, and generative adversarial networks for video generation.

2. Literature Review

AI's prowess is not confined to manipulating existing faces. It can conjure up faces that do not exist in reality, making it an invaluable asset for content creators.

This study delves into the realm of AI-generated faces and voices. Our investigation focuses on whether human observers can discern this synthesized content from real footage, when produced by amateurs using readily available AI tools.

In this literature review section, we will lay down a comprehensive framework for creating entirely artificial content by amateurs. Additionally, we will delve into an in-depth analysis of the tools currently available in this rapidly evolving field

2.1. Artificial Face Generation

The literature on artificial face generation has grown rapidly, with StyleGAN emerging as the frontrunner in the field. [KLA19a, KLA*20] from NVIDIA first proposed a style-based generative adversarial network that improved traditional distribution quality metrics and disentangled latent factors of variation.

Subsequent improvements to StyleGAN include generator normalization, progressive growing, and generator regularization for better latent code-to-image mapping [KLA19b]. These enhancements enabled the generation of not only more realistic human faces but also stylistically diverse faces, such as anime characters [Bra19].

However, variations of StyleGAN have been developed, such as TediGAN, which enables text-based editing of facial attributes. Different art styles, such as cartoonish or hyper-realistic, can also be chosen depending on the desired outcome.

Other options include LiftedGAN and StyleRig, which transform two-dimensional faces into three-dimensional ones. For our project, we opted for the basic StyleGAN, as it met our goal of generating a convincing synthetic face without specifying particular attributes.

Although generative adversarial networks (GANs) dominate the field, alternative methods exist. Autoencoders, for instance, have recently been upgraded with the development of the Adversarial Latent Autoencoder (ALAE) and StyleALAE. These autoencoders generate high-quality face images and enable face reconstructions and manipulations based on real images [PAD20].

Advancements in StyleGAN have led to auxiliary applications, such as StyleRig, which enables face rig-like control over generated portraits [TEB*20], and TediGAN, which uses text to manipulate generated images [**?**]. LiftedGAN, another extension of StyleGAN, generates images and their 3D components by distilling prior knowledge from StyleGAN2 [SAJ21]. Despite recent improvements, challenges remain. For example, editing specific attributes of generated faces sometimes inadvertently alters other features [KGM*22].

2.2. Audio Synthesis

Audio synthesis has been an area of interest for researchers, as voices provide crucial information about a speaker's identity and characteristics [MJTDF95]. Early synthetic voices were used in Augmentative and Alternative Communication (AAC) devices to help those who could not speak [BPYG05]. However, initial synthetic voices were less credible compared to real ones [CORP06], and the uncanny valley effect emerged when the realism of a character's face and voice did not match [MSL*11].

Advancements in synthetic voices have made them more viable when matched correctly with visuals [CCZM17]. Voice cloning, which uses reference audio to synthesize speech, has also progressed [JZW*18]. Expressing emotions in synthetic voices remains a challenge, as it is vital for the voice to match the generated visuals [ZX20].

Voice synthesis is commonly achieved through voice cloning. Software like Descript streamlines this process by requiring users to record provided lines, after which the software generates the desired voice. However, this approach still necessitates the involvement of a voice actor. Alternatively, Descript offers a library of pre-generated voices that can be used for various projects.

2.3. Deepfake Creation and Detection

Deepfake technology creates realistic manipulations of videos, posing challenges for detection. Early detection methods relied on artifacts left by the synthesis process [LL18], but as new manipulation methods emerge, detection techniques that require minimal training data are needed [RCV*19]. Integrating detection methods into distribution platforms can increase their effectiveness [NNN*19].

The First Order Motion Model enables animation of 2D images based on key points and local affine transformations [SLT*19], while DeepFaceLab provides an accessible and adaptable face-swapping platform [PGC*20].

To create a cohesive deepfake, we evaluated different tools, including DeepFaceLab and the First Order Motion Model. The latter was chosen due to its simplicity and accessibility to those without prior knowledge in the field.

2.4. Video Generation

The domain of video generation, which inherently includes the generation of images, has emerged as a significant research focus. Investigations have delved into the production of images and videos through textual prompts. Techniques such as the Zero-Shot Text-to-Image Generation [RPG*21] manifest images from textual descriptions. Conversely, the TiVGAN approach [KJK20] facilitates the generation of comprehensive videos from text, constructing each frame as individual images. These methods provide the capability to fabricate entirely synthetic establishing scenes or backgrounds for media content.

Despite these advancements in video generation, there remain limitations that hinder greater control and manipulation in creating synthetic media. Although the current techniques excel at producing visually engaging content, they still lack the finesse to integrate more intricate elements, particularly in the realms of audio and script generation. Combining the individual functionalities of various tools presents an intriguing avenue for exploration and thus was what we decided to investigate in our framework.

3. Research Questions and Hypotheses

Our study aims to investigate the credibility of synthesized faces and voices created by amateurs using artificial intelligence tools. Based on the literature review, we propose the following research questions:

1. Can amateurs (just introduced to the tools) generate convincing synthesized faces and voices using available AI tools?
2. How do humans perceive the realism of these synthesized faces and voices when combined in video content?
3. What factors influence the credibility of the synthesized faces and voices?

Based on the advancements in AI-generated faces and voices, we hypothesize that:

1. Amateurs will be able to generate realistic synthesized faces and voices using AI tools.
2. Humans will struggle to differentiate between real and synthesized faces and voices when combined in video content.
3. The credibility of the synthesized faces and voices will be influenced by the quality of the generated visuals, the emotional expressiveness of the synthetic voices, and the alignment between the visuals and the audio.

In this paper, we will present research conducted to test these hypotheses and answer the research questions. The results will contribute to our understanding of the implications and potential applications of AI-generated media.

4. Method and Implementation

We tested the viability and ease of creating artificially generated media by having 2 amateurs (people who had 0 hours with any of the software and were unfamiliar with the process) produce the artifacts and evaluating their quality.

Our project required the generation of both visuals and audio by amateurs. We used StyleGAN for face generation, either by running the publicly available code or by accessing websites like thispersondoesnotexist.com. This approach makes face generation accessible to individuals without coding experience. In our project, multiple faces were generated using StyleGAN. Finally, we used Descript for voice synthesis.

After generating faces with StyleGAN and voices with Descript, we produced a script for the artifacts. Videos of people reciting the script were recorded, and the First Order Motion Model was used to replace the real faces with the synthesized ones. The generated voices were then synchronized with the new visuals.

We conducted a survey to evaluate participants' ability to differentiate between the realism of videos created by amateurs, specifically focusing on their capability to recognize the artificial nature of the media and pinpoint aspects that appeared to be synthesized.

4.1. Design and Development

To create the synthesized face, we utilized StyleGAN, either by running the publicly available code or by accessing websites like thispersondoesnotexist.com. This approach makes face generation accessible to individuals without coding experience. In our project, multiple faces were generated using StyleGAN, as shown below.

Upon obtaining the desired faces generated by StyleGAN, we proceeded to create artificial voices to accompany them. We employed Descript, a widely accessible and free-to-use software, offering higher-tier paid options with an expanded library of artificially generated voices and enhanced support. Descript allows users to record a person's voice and create an artificial voice for text-to-speech applications. The primary challenge lay in identifying voices that complemented the generated faces, either by enlisting multiple individuals with suitable voices or utilizing the pre-existing artificial voices in Descript's library.

For the project's purposes, the basic free version of Descript sufficed, and its library of available artificial voices provided ample options to match the generated faces.

We created a pipeline (fig 1) for generating artificial content. First, we selected artificial voices and used AI to generate sentences

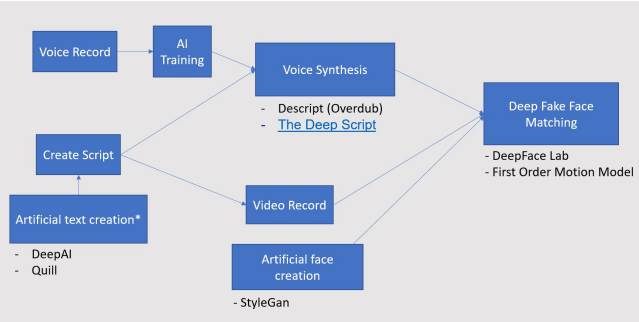

Figure 1: *Proposed pipeline followed by the amateurs to generate the videos*

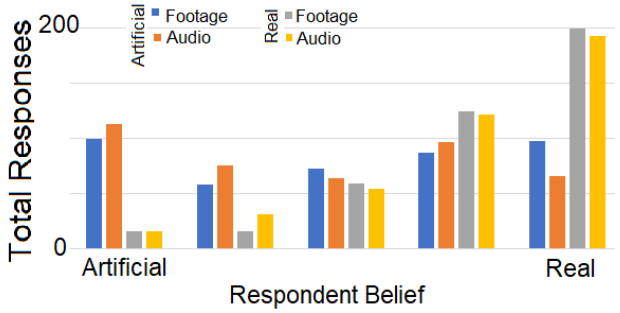

Figure 2: *Summation of the results, presented of the 16 total videos, divided by Footage and Audio, and Artificial and Real, bars represent the number of people who answered, how believable each of the components were*

for both real and fake media, ensuring equal spoken content. Finally, we used the First Order Motion Model and Google Colab to create deepfakes by combining real footage with artificially generated faces. This approach provided accessible and user-friendly deepfake creation, even for those without high-end hardware.

After creating the deepfakes, the next step was to merge audio and visual components using video editing software, such as Premiere Pro. This process involved meticulous synchronization of audio and visual elements, which may require adjustments to either component. Although other video editing software could have been used, the choice of Premiere Pro was based on the authors' familiarity and prior experience.

5. Evaluation

To evaluate the realism of the fully artificial artifacts, we designed a survey incorporating both genuine and deepfake videos. We recorded videos of individuals reciting the AI-generated sentences to serve as facial data for deepfake creation and genuine artifacts for comparison.

We collected data using Amazon Mechanical Turk, a crowdsourcing platform. Participants were recruited from a diverse range of ages (18-56 $\mu = 32$ $\sigma = 2$) and backgrounds. Prior to completing the survey, participants were asked to report any mental or visual issues that might affect their perception of stimuli.

In total, 60 participants (32 Female) completed the survey. We excluded 8 participants due to unreliability in their responses. The final dataset consisted of 52 participants.

The survey included stimuli of both faces and voices, which were rated on a 5-point scale ranging from "For Sure Artificial" to "For Sure Real". The rating scale was used for both the face and voice stimuli separately to maintain consistency across the survey.

Additionally, participants could provide feedback on specific aspects they believed were genuine or fake within each video.

5.1. Presented Content

To assess the perceived authenticity of the generated deepfake media, eight different deepfakes were created using the methodology outlined in previous sections. These deepfakes were randomly presented along side eight genuine videos featuring individuals uttering artificially generated sentences (to maintain fairness in terms of the spoken content's context).

6. Results and Discussion

The anonymous online survey collected responses from fifty-two participants who evaluated the authenticity of the visuals and audio for each of the artificial and genuine media artifacts. The results are presented in fig 2 with a rating of one indicating the media is perceived as fake and five indicating it is perceived as real.

Participants had varying perceptions of the authenticity of the visuals and audio in each artificial video. In the first video, the majority perceived the visuals to be fake, while the audio was perceived as more authentic. In the second video, opinions on the visuals were divided, but the audio was generally perceived as fake. The third video received mixed ratings, with insufficient mouth movement cited as an issue. The fourth video had mixed ratings for both visuals and audio, with more participants believing the visuals were real. In the fifth video, most participants believed the visuals were real, while opinions on the audio were divided. The majority of participants in the sixth video believed the visuals were real, but the audio was perceived as fake. The seventh video received lower authenticity ratings for both the audio and visuals. The eighth video was generally rated as low quality, with low authenticity ratings for both audio and visuals.

There was a strong, positive correlation between the evaluators as a whole believing the footage and the audio were artificial (the average of their scores), which was statistically significant ($b = .618$, $p < .034$) when considering only the artificial video and ($b = .590$, $p < .003$) when considering the two sets.

Overall, participants consistently perceived the real videos and audios as authentic. However, for the fake videos, ratings were more variable. In particular, the audio was more often perceived as fake than the visuals. In the qualitative responses, the most common comment was that the audio did not match the appearance of the face, and they further commented that one way they used

to differentiate was the "attractiveness" of the person in the video, meaning the real one had less attractive people than the artificially generated one.

This suggests that participants may have relied heavily on visual cues when assessing the authenticity of the stimuli.

Overall, these findings highlight the importance of ensuring that audio and visual components are properly synchronized and matched when creating artificial stimuli, in order to enhance their perceived authenticity.

The successful creation of multiple artificial media artifacts demonstrates the plausibility of generating entirely artificial content. However, the quality of these artifacts plays a crucial role in their perceived authenticity. Creating low-quality, obviously, artificial media is relatively easy, whereas generating convincing content that could deceive viewers requires greater effort.

Our results suggest, that given a face distinguishing artificial audio from real audio is relatively easier. Artificial visuals, on the other hand, seem to generate more ambiguous perceptions. Numerous factors could contribute to participants recognizing fake audio, such as poor synchronization between visuals and audio, or mismatched voice characteristics and facial appearances.

Other considerations include pronunciation, tone, and accent in relation to the speaker's appearance and mouth movements. Inconsistencies in audio and visual quality may also influence perceptions of authenticity. Furthermore, since the survey explicitly asks about fakeness, participants might scrutinize content more closely than if they were unaware of potential artificial elements.

7. Conclusions and future work

In this work, we investigated the credibility of a series of deepfake videos created by amateurs from our school. We presented these artificial videos, along with baseline real videos, to participants on Amazon MTurk and asked them to rate their authenticity. It is important to note that in this context, the artificial videos were completely computer-generated, including the audio.

Our results show that there was a general trend toward perceiving the videos as fake, with the majority of participants rating them as less authentic than the real videos. However, we did observe variations in authenticity ratings across the different videos. We found that participants were more likely to perceive the visuals as fake when there was a mismatch between voice and face or when there was insufficient mouth movement for pronouncing words. Meanwhile, participants were more likely to perceive the audio as fake when there was a mismatch between the voice and the person in the video.

It is important to note that this study was performed in 2021. The year in which the study is conducted can have implications for the generalizability and relevance of the findings, particularly in fast-moving fields such as deepfake technology. In the case of our study, the findings should be interpreted in the context of the state of the technology in 2021. As deepfake technology continues to evolve rapidly, future research will need to replicate and extend our findings to keep pace with these developments. Additionally, our study was conducted during a time of heightened public awareness and concern about the potential harms of deepfakes, which may have influenced our participants' perceptions of the videos.

Even though the respondents were unaware that the audio and video were always matching whether artificial or not, it is still an interesting and valid research question to examine if the combination of one artificial component with one real component changes the results or level of confusion.

Nevertheless, based on our findings, we recommend that future investigations should focus on generating audio that matches one's appearance more accurately. Additionally, we suggest that more attention be paid to the creation of "average" looking artificial people, as this may increase the credibility of the videos.

Overall, our study contributes to the understanding of how the credibility of deepfake videos is perceived by viewers. It highlights the importance of considering both the visual and audio components of these videos, and provides insights that can inform the development of more realistic deepfakes in the future.

References

[BPYG05] BUNNELL H. T., PENNINGTON C., YARRINGTON D., GRAY J.: Automatic personal synthetic voice construction. In *Ninth European Conference on Speech Communication and Technology* (2005). 2

[Bra19] BRANWEN G.: Making anime faces with stylegan. https://www.gwern.net/Faces, 2019. 2

[CCZM17] CABRAL J. P., COWAN B. R., ZIBREK K., MCDONNELL R.: The influence of synthetic voice on the evaluation of a virtual character. In *INTERSPEECH* (Stockholm, 2017), pp. 229–233. 2

[CORP06] CABRAL J., OLIVEIRA L., RAIMUNDO G., PAIVA A.: What voice do we expect from a synthetic character? In *Proceedings of SPECOM* (2006), Citeseer, pp. 536–539. 2

[JZW*18] JIA Y., ZHANG Y., WEISS R., WANG Q., SHEN J., REN F., NGUYEN P., PANG R., MORENO I. L., WU Y.: Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems* (2018), vol. 31. 2

[KGM*22] KHODADADEH S., GHADAR S., MOTIIAN S., LIN W.-A., BÖLÖNI L., KALAROT R.: Latent to latent: A learned mapper for identity preserving editing of multiple face attributes in stylegan-generated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022), pp. 3184–3192. 2

[KJK20] KIM D., JOO D., KIM J.: Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access 8* (2020), 153113–153122. 3

[KLA19a] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 4401–4410. 2

[KLA19b] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 4401–4410. 2

[KLA*20] KARRAS T., LAINE S., AITTALA M., HELLSTEN J., LEHTINEN J., AILA T.: Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 8110–8119. 2

[LL18] LI Y., LYU S.: Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656* (2018). 3

[MJTDF95] MULLENNIX J. W., JOHNSON K. A., TOPCU-DURGUN M., FARNSWORTH L. M.: The perceptual representation of voice gender. *The Journal of the Acoustical Society of America 98*, 6 (1995), 3080–3095. 2

[MSL*11] MITCHELL W. J., SR K. A. S., LU A. S., SCHERMERHORN P. W., SCHEUTZ M., MACDORMAN K. F.: A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception* 2, 1 (2011), 10–12. 2

[NNN*19] NGUYEN T. T., NGUYEN C. M., NGUYEN D. T., NGUYEN D. T., NAHAVANDI S.: Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573* (2019). 3

[PAD20] PIDHORSKYI S., ADJEROH D. A., DORETTO G.: Adversarial latent autoencoders. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 14104–14113. 2

[PGC*20] PEROV I., GAO D., CHERVONIY N., LIU K., MARANGONDA S., UMÉ C., DPFKS M., FACENHEIM C. S., RP L., JIANG J., ET AL.: Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535* (2020). 3

[RCV*19] ROSSLER A., COZZOLINO D., VERDOLIVA L., RIESS C., THIES J., NIESSNER M.: Faceforensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), 1–11. 3

[RPG*21] RAMESH A., PAVLOV M., GOH G., GRAY S., VOSS C., RADFORD A., CHEN M., SUTSKEVER I.: Zero-shot text-to-image generation. *International Conference on Machine Learning* (2021), 8821–8831. 3

[SAJ21] SHI Y., AGGARWAL D., JAIN A. K.: Lifting 2d stylegan for 3d-aware face generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 6258–6266. 2

[SLT*19] SIAROHIN A., LATHUILIÈRE S., TULYAKOV S., RICCI E., SEBE N.: First order motion model for image animation. *Advances in Neural Information Processing Systems 32* (2019). 3

[TEB*20] TEWARI A., ELGHARIB M., BHARAJ G., BERNARD F., SEIDEL H.-P., PÉREZ P., ZOLLHOFER M., THEOBALT C.: Stylerig: Rigging stylegan for 3d control over portrait images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 6142–6151. 2

[ZX20] ZHU X., XUE L.: Building a controllable expressive speech synthesis system with multiple emotion strengths. *Cognitive Systems Research 59* (2020), 151–159. 2